



# Big Data Analytics for Prostate Radiotherapy

James Coates<sup>1\*</sup>, Luis Souhami<sup>2</sup> and Issam El Naqa<sup>3</sup>

<sup>1</sup> Department of Oncology, University of Oxford, Oxford, UK, <sup>2</sup> Division of Radiation Oncology, McGill University Health Centre, Montreal, QC, Canada, <sup>3</sup> Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA

Radiation therapy is a first-line treatment option for localized prostate cancer and radiation-induced normal tissue damage are often the main limiting factor for modern radiotherapy regimens. Conversely, under-dosing of target volumes in an attempt to spare adjacent healthy tissues limits the likelihood of achieving local, long-term control. Thus, the ability to generate personalized data-driven risk profiles for radiotherapy outcomes would provide valuable prognostic information to help guide both clinicians and patients alike. Big data applied to radiation oncology promises to deliver better understanding of outcomes by harvesting and integrating heterogeneous data types, including patient-specific clinical parameters, treatment-related dose–volume metrics, and biological risk factors. When taken together, such variables make up the basis for a multi-dimensional space (the “RadoncSpace”) in which the presented modeling techniques search in order to identify significant predictors. Herein, we review outcome modeling and big data-mining techniques for both tumor control and radiotherapy-induced normal tissue effects. We apply many of the presented modeling approaches onto a cohort of hypofractionated prostate cancer patients taking into account different data types and a large heterogeneous mix of physical and biological parameters. Cross-validation techniques are also reviewed for the refinement of the proposed framework architecture and checking individual model performance. We conclude by considering advanced modeling techniques that borrow concepts from big data analytics, such as machine learning and artificial intelligence, before discussing the potential future impact of systems radiobiology approaches.

**Keywords:** radiotherapy, data mining, machine learning, big data, systems radiobiology

## INTRODUCTION

Prostate cancer is the second most common cancer among men and is the fourth most common cancer overall (1). In Europe alone, prostate cancer is the most commonly diagnosed cancer in men and accounts for approximately one-quarter of newly diagnosed cases per annum (2).

Fractionated radiation therapy (radiotherapy) is a primary treatment method for prostate cancer patients with localized disease – approximately one-quarter of patients have some form of radiotherapy incorporated into their treatment regimen (3). The widespread acceptance of radiotherapy as a first-line treatment modality can be attributed to high rates of local control and acceptable levels of normal tissue toxicity (4, 5).

Modern external beam radiation therapy (EBRT) delivery technologies, such as stereotactic body radiation therapy (SBRT) and volume-modulated arc therapy (VMAT), offer increased conformity and total dose while minimally damaging adjacent normal structures (6–8).

## OPEN ACCESS

### Edited by:

Adam Paul Dicker,  
Thomas Jefferson University, USA

### Reviewed by:

Yaacov Lawrence,  
Sheba Medical Center, Israel  
Matthew T. Studenski,  
University of Miami, USA

### \*Correspondence:

James Coates  
james.coates@oncology.ox.ac.uk

### Specialty section:

This article was submitted to  
Radiation Oncology, a section of the  
journal *Frontiers in Oncology*

**Received:** 29 March 2016

**Accepted:** 31 May 2016

**Published:** 14 June 2016

### Citation:

Coates J, Souhami L and El Naqa I  
(2016) Big Data Analytics for Prostate  
Radiotherapy.  
*Front. Oncol.* 6:149.  
doi: 10.3389/fonc.2016.00149

These advanced treatment tools generate vastly more amounts of treatment-related data than contemporary counterparts, such as three-dimensional conformal radiation therapy (3D-CRT). In terms of outcomes analysis, this can render quantitative modeling of treatment plans and retrospective outcomes exploration more complicated.

Historically, dose–volume metrics alone were used in an attempt to explain aberrant toxicities or biochemical relapses (9). Canonical examples of this include either hot spots in overlapping regions between PTV and normal structures that were thought to independently induce adverse normal tissue effects or, conversely, suboptimal PTV coverage thought to be the main cause for inadequate local control (10). In recent years, however, it has been demonstrated that dose–volume metrics, while straightforward to obtain and contributing significantly to the analysis of radiotherapy outcomes, are not the only determining factors of success in predicting radiotherapy outcomes (11, 12). This has been shown in prospective application of dose–volume metrics whereby such metrics have proven to provide limited classification performance (13, 14). Aside from dose–volume data, the emergence of advanced imaging modalities and high-throughput “-omics” methods have led to the generation of enormous amounts of data that can similarly be used to predict outcomes.

The multi-dimensional space that EBRT-related biological, dosimetric and clinical variables span is referred to herein as the *RadoncSpace*. Two overarching predictive modeling approaches that exploit big datasets and search different sub-spaces of the *RadoncSpace* have surfaced in recent years: *radiomics* (use of imaging datasets for outcome prediction) (15) and *radiogenomics* (uncovering relationships between biological data and outcomes) (16). Pioneering application of these two techniques speaks to the ever-increasing application of data-mining techniques and big data analytics (the so-called “*panomics*”) to modern oncology (4).

The specific objective of big data analytics in radiotherapy is to develop predictive models that capture underlying factors contributing to the development of selected endpoints without over-fitting noise or under-fitting trends. In line with the nature of big data and the heterogeneity of patient populations, a strict requirement of such modeling frameworks is that input datasets must be large enough to include variability, which accurately reflects the underlying patient population. Otherwise, resulting models can suffer from poor prospective prediction performance.

Clinically, it has already been demonstrated how such models could be used to better inform patients of treatment-associated risks. Namely, by integrating outcome models into treatment planning systems (TPSs) and recommending dose-escalation or dose-reduction (11). Given the potential future impact of outcome models in the clinic, the selection of tools and models for the fabrication of a predictive framework must be chosen carefully in a way to facilitate identification of optimal models.

In this work, we focus our attention exclusively on outcomes associated with EBRT; however, the presented modeling techniques are easily generalizable to any dose distribution. In this work, we briefly review the radiobiology of prostate cancer as a basis for understanding the theoretical underpinnings of analytical outcome models. Analytical models attempt to predict radiation-related toxicities by formalizing abridged versions of the

biological processes by which selected endpoints become manifest. Subsequently, we discuss big data and data-driven modeling approaches based on techniques previously used successfully for exploring outcomes in radiotherapy. In contrast to analytical models, data-driven models are entirely empirical in nature, potentially making them more robust albeit more difficult to analyze or interpret. We then consider techniques that optimize model parameters in order to maximize model robustness and prevent under-/over-fitting, which are two common pitfalls in big data outcome modeling. The article concludes by presenting modeling techniques based on advanced artificial intelligence as well as on systems theory.

## PROSTATE CANCER

### Pathology

Adenocarcinoma of the prostate is the most common histopathological type of prostate malignancy and typically arises in the peripheral zone (17). Up to one-half of men present with prostate cancer at time of autopsy although tumors identified in many of these cases are typically small, impalpable, and of low grade (18).

Prostate tumors are known to have remarkable biological heterogeneity from patient to patient and even across tumor volumes (19–21). The metastatic potential of prostate cancer is similarly variable and is furthermore reflected in the wide variation in overall survival rates for those with localized disease at time of diagnosis (21). Notably, the high degree of heterogeneity makes standardization and characterization of prostate adenocarcinoma phenotypes challenging and institution specific.

### Basic Radiobiology

The  $\alpha/\beta$  originates from the linear-quadratic (LQ) formulation of *in vitro* cellular survival experiments (Eq. 1):

$$SF = e^{-(\alpha D + \beta D^2)} \quad (1)$$

where SF is the surviving fraction after a dose (D). The coefficient  $\alpha$  [ $\text{Gy}^{-1}$ ] in front of the linear dose term (D) relates to single-hit inactivation and the  $\beta$  [ $\text{Gy}^{-2}$ ] coefficient pertains to the expected rate of double-hit (two-track) cellular inactivation (22, 23). The  $\alpha/\beta$  ratio taken from the LQ model allowed numerous fundamental radiobiological questions to be answered quantitatively. It remains a relevant parameter in radiotherapy today due to its clinical significance as a measure of tissue-specific fractionation-sensitivity.

It is well known that prostate cancers are relatively slow-growing malignancies with low  $\alpha/\beta$  ratios (24–27), unlike most malignancies. When used in the context of biologically effective dose (BED) (Eq. 2), the low prostate  $\alpha/\beta$  ratio (~1.5 Gy) translates to a high sensitivity to fraction size:

$$BED = D \cdot \left( 1 + \frac{d}{\alpha/\beta} \right) \quad (2)$$

where *D* is the total dose of the radiotherapy regimen and *d* is the fraction size. Furthermore, since the  $\alpha/\beta$  ratio for prostate cancer

is lower than that of normal tissues ( $\sim 3$  Gy), an improved therapeutic ratio can be expected using hypofractionation (28–30).

In 2013, Vogelius and Bentzen performed a meta-analysis of 1965 patients derived from five separate studies (31). In line with pioneering work by Fowler and colleagues (25), they showed that prostate cancers do indeed have an exceedingly low  $\alpha/\beta$  ratio. Interestingly, after accounting for changes during treatment, their estimate of  $\alpha/\beta$  increased. This may indicate that the  $\alpha/\beta$  for prostate tumors changes throughout the course of radiotherapy treatment, probably due to subpopulation selected induced by radiotherapy itself.

Although  $\alpha/\beta$  has provided insight into radiobiology of prostate cancer, it remains unclear how relevant the ratio is in cases of modern EBRT delivery, such as high-dose hypofractionated SBRT regimens, mixed-modality treatments (photon with proton boost) or when using charged particles, such as carbon ion, all of which are becoming increasingly popular treatment options for prostate cancer. In such cases, aggregation of large-scale datasets serving as inputs to big data analytics may provide more useful insight either as a supplement or as a substitute to classical paradigms in radiobiological modeling.

## TYPES OF OUTCOMES

Toxicity outcomes in radiotherapy can be segregated into two categories: acute (effects observed within 3 months after the termination of radiotherapy) and late (effects that manifest after the 90-day cutoff). Furthermore, normal tissue damage can be segregated by site; in prostate radiotherapy, normal tissue side effects manifest themselves most frequently as one or more of gastrointestinal (GI) toxicities, genitourinary (GU) toxicities, or erectile dysfunction (ED).

### Acute (Early) Outcomes

Acute effects due to normal tissue damage from ionizing radiation in prostate cancer radiotherapy regimens include GI/GU symptoms. Acute symptoms are most often transient, self-limiting events in that they appear and resolve within a matter of weeks without contributing significantly to severe or long-term morbidity, although some consequential late effects in prostate radiotherapy have indeed been recorded (32–34).

A 2015 review article by Drodge et al. compiled the results of 22 prospective hypofractionated trials completed between 2001 and 2013 (35). Using the RTOG/EORTC toxicity grading scheme and including studies that used different treatment modalities and schedules, the authors concluded that Grade 3 acute toxicities, on the whole, affect less than 10% of prostate patient cohorts receiving hypofractionated EBRT. Furthermore, Grade 2 toxicities affected under half of patients. The study also expanded upon the practical challenges in interpreting outcome data from independent trials that may use different grading schemes or endpoints.

Proton therapy is becoming more common in modern times for use in treating prostate cancer (36). Studies have shown that the frequency of acute effects with proton for prostate cancer is not significantly increased over that of conventionally fractionated photon therapy regimens (37, 38). One dose-escalation study with 85 prostate cancer patients using proton doses up to

82 Gy-equivalent (GyE) yielded acute toxicity levels comparable to photon radiation (39).

## Late Normal Tissue Endpoints

Radiation-induced late normal tissue damage consists of toxicities that occur  $>90$  days after completion of radiotherapy. Late toxicities can range from mild, moderate, severe to life-threatening requiring immediate intervention. They are categorized as either GI or GU effects and many retrospective studies report only these outcomes; however, sexual dysfunction is also considered herein. Unfortunately, the full pathophysiology of the radiation-induced manifestation of ED has yet to be fully elucidated.

The difficulty in assessing late toxicities is that often times no quantitative physiological evidence exists or can readily be obtained. Grading schemes have been developed to resolve such issues. Grading schemes require physicians to assign integer values to the radiation-induced side effect based on selected criteria. Some schemes utilize self-scoring questionnaires, while others rely on grades assigned by attending oncologists. Interestingly, several groups have sought to explore the correlation between different scoring schemes using a single set of data in order to explore what role grading schemes have on incidence rates of toxicity (40–42). Such works bear significance for the use of big data analytics as many frameworks utilize supervised learning techniques that rely on the accuracy of outcome measures.

## Local Control Endpoints

It is estimated that overall approximately one-third of prostate cancer patients experience some type of biochemical relapse within the first decade after completion of their EBRT treatment regimen (43). In reporting local control outcomes, clinical studies typically do so according to specific criteria, such as the ASTRO-RTOG Phoenix definition of local biochemical failure (44). Guidelines often include prostate serum antigen (PSA) scores although derivatives of simple PSA scores have also been considered, such as PSA doubling time (43). However, it should be noted that a rising PSA does not always indicate a local failure and it can antedate the diagnosis of metastatic disease for many years; thus, caution should be exercised when using it as a surrogate for local failure endpoint.

## DATA TYPES

### Dose–Volume Metrics

Typical dose–volume metrics used in outcomes modeling include dose to a given volume or volume of tissue receiving at least a particular dose. These parameters can be readily extracted from dose–volume histograms at the treatment planning stage. Physiological changes, such as weight gain/loss or changes in tumor composition or anatomical position, may take place during treatment and, thus, dose delivered may not necessarily reflect biologically absorbed dose. It is likely that dose–volume variables could have their predictive accuracy improved by incorporating intra-fractional computed tomography (CT) scan changes, as has been considered in literature (45).

The equivalent uniform dose (EUD) (46) is a dose–volume metric that can be used to describe inhomogeneous dose

distributions. The generalized EUD (gEUD) (47) is a further extension used for normal tissues of interest (Eq. 3):

$$\text{gEUD} = \left( \sum_i v_i D_i^a \right)^{1/a} \quad (3)$$

where the variables  $v_i$  is fractional volume for the tissue exposed to dose  $D_i$ , the parameter  $a$  is a factor relating to the volume effect of a given tissue type. These two metrics appear oftentimes in analytical outcome models as they serve as excellent tools to summarize dose distributions across volumes.

## Clinical Parameters

Clinical data can be parameterized and used to investigate covariates of interest. An example in the case of prostate cancer is in patients receiving anti-coagulant therapy and presenting with rectal bleeding (RB) late in their follow-up period, which can otherwise easily be mistaken for a late toxicity. Another case is the combined use of androgen deprivation therapy (ADT) since ED can be a side effect of ADT alone and thereby lead to an increased prevalence of late ED in a given prostate patient cohort.

## Spatial Parameters

Spatial dose–volume histograms (zDVHs) can be used to compare spatial treatment planning information to outcomes (48–50). The advantage of incorporating spatial information is that it provides modeling frameworks information about the location of dose extremes and, thus, mitigates having to rely solely on approaches based on volume-averages (or gEUD). This reduces the risk of under- or over-valuing the contribution of hot or cold spots. Spatial data can also provide information related to the contribution of hot spots in accessory structures, for example, in the case of rectal contour overlap with the PTV.

## Biological Variables

Several types of biological variables have been used previously in attempting to elucidate mechanisms by which prostate radiotherapy toxicities become manifest. The most popular class of variables found in literature today is related to genetic mutations. Additionally, work has been performed on exploring the role of epigenetics (51) and transcript expression levels (52) in long-term radiotherapy outcomes.

## Genetic Variables

Given the relatively disappointing prospective predictive power of singular classes of genetic variables on their own (53–56), it is likely that modeling frameworks will need to allow for several types to be incorporated in a given model in order to maximize prospective classification performance.

Single-nucleotide polymorphisms (SNPs) consist of single-nucleotide changes. Their presence in certain genes or regulatory regions has been shown to be well correlated with prostate radiotherapy-related outcomes (52, 54, 57, 58). This is probably due to altering functional transcripts or protein confirmations after translation.

Further to SNPs, copy number variations (CNVs) have recently been of increasing interest to the radiotherapy community (59).

CNVs reflect the number of copies of a particular gene and are, therefore, larger structural genetic mutations than SNPs. This could mean that larger changes in a given genome could be seen with CNV changes.

In our previous work, we have shown the value of integrating CNVs alongside SNPs (of the same gene) together with dose–volume metrics (60). Specifically, we have demonstrated that changes in the gene concentration of DNA repair gene XRCC2 can predict severe (Grade 3) late RB for hypofractionated prostate patients treated with 3D-CRT. More importantly, the resulting radiogenomic models led to increased predictive power as compared to using either type of genetic variable alone. We, furthermore, demonstrated that the improvement using SNPs and CNVs is not limited to data-driven frameworks but could also be applied to analytical models. These results indicate that different genetic mutations in the same gene may contribute similarly to a given outcome. If proven to be the case, it is likely a result of outcome scores being limited snapshots of complex pathophysiological events reflecting more than one biological alteration.

## Integrating Genetic Variables in Outcome Models

In the case of data-driven modeling, genetic parameters can be considered as independent variables and regressed alongside clinical risk factors and dose–volume metrics. For analytical models, the method in which genetic parameters are integrated depends on the nature of the model at hand. In 2006, two groups showcased how dose-modifying factors (DMFs) extracted from clinical risk factors could be used to stratify standard analytical models and thereby generate “mixed” data-type models (57, 61). In 2013, Tucker et al. expanded this approach to include SNPs using an approach easily generalizable to any biological variable and demonstrated significantly improved classification performance (62). Rancati and colleagues further extended this approach using clinical risk factors for Logit and EUD models (63), from which our group drew inspiration in developing radiogenomic models using biological, clinical, and dosimetric variables (60).

## Other Biological Variables

### Epigenetics

The importance of epigenetic alterations to the genetic code has not by any means been understated by the scientific community in recent years (64–67). However, the significance of epigenetic modifications in radiotherapy remains to be fully understood. Research related to epigenetics and radiotherapy could be complicated by the fact that mounting evidence implies that radiotherapy itself can induce epigenetic changes (67).

Thus far, thousands of differentially methylated regulators have been identified in many cancer types thanks to epigenome-wide association studies (EWAS) (68). Differentially regulated promoter may serve as novel biomarkers to predict risk of biochemical relapse or serve as indicators of normal tissue radiosensitivity. In prostate cancer specifically, wide-ranging hypo- and hyper-methylations have been identified that correlated with early-stage carcinogenesis and aggressive tumor phenotypes (69, 70). Efforts are underway to generate an epigenetic code (71–73), which may facilitate the ability to perform and interpret EWAS

results as well as provide a new class of input data for outcome models (74).

### **High-throughput Proteomics and mRNA Expression Levels**

Numerous methods used to quantify large numbers of biological factors have been pioneered and introduced into mainstream biology research within the last decade. These technologies include well-characterized microarrays and proteomic analysis technologies that can quantify the levels of expression of up to tens of thousands mRNA transcripts or proteins in a single sample.

After generating large quantities of data, high-throughput modeling frameworks can be used that are able to deal with large numbers of variables (75). This approach has been used successfully in clinical oncology to stratify tumor phenotypes and estimate prognoses to help guide optimal therapeutic regimens (76–78). In the case of radiation oncology, high-throughput data have yielded several multi-gene signatures for hypoxia (79–81).

The challenges of utilizing a large number of variables in outcome models are well summarized by the multiple testing dilemma: too few samples relative to a large number of variables being tested can lead to spurious correlations. Even after utilizing simple supervised learning algorithms to pre-process the data, the number of mRNA transcripts that a single microarray experiment can yield is often in the thousands (82). This issue can be mitigated by large-scale validation studies but these are expensive, time-consuming and patient accrual can limit achieving the necessary sample size.

Alternatively, methods in artificial intelligence are becoming increasingly popular to explore the complex, hidden relationships between outcomes and biological variables (83). In contrast to brute-force estimating of correlations, machine-learning techniques in artificial intelligence have the ability to process highly structured, high-dimensional data while controlling for over- and under-fitting by drawing on methods from control, probability, and information theory.

## **MODELING TECHNIQUES**

### **Risk Quantification**

The likelihood of obtaining local control is quantified mathematically by tumor control probability (TCP). TCP is a probability that indicates chances for success of a treatment according to a particular endpoint, usually long-term control. Many studies using TCP-based approaches have shown that cancer cells *in situ* have complex, high-dimensional repopulation kinetics when exposed to ionizing radiation and/or chemotherapy (84–87). Such kinetics can lead to complex models and be dependent upon several factors, such as repair capacity, quality of radiation, fractionation scheme, and surrounding microenvironment (88).

In the case of normal tissue side effects from radiotherapy, risk is quantified via normal tissue complication probability (NTCP). NTCP values can be tailored to each individual treatment regimen to reflect the risk of a given side effect. Conventionally, such frameworks were limited to dosimetric data; however, it is now understood that late normal tissue toxicities are furthermore functions of a variety of biological, physical, and clinical factors (16, 89).

### **Analytical Modeling**

As previously described, models of the analytical class are based on simplified theoretical mechanisms of action radiobiological interactions. They include some level of mechanistic insight into a specific mechanism by which radiotherapy outcomes become manifest and are, therefore, also referred to as *mechanistic models*.

### **Tumor Control Probability**

Cells that can lead to tumor growth are termed tumorigenic stem cells or cancer stem cells. These cells are, in theory, the primary targets of anti-cancer therapies. The probability that a given treatment will induce eradication of cancer stem cells for a given patient is mathematically given by the TCP.

### **The Linear-Quadratic**

The LQ model has gained popularity in literature since it follows survival closely at conventional doses of radiation. Furthermore, the model provides a simplified theoretical basis for how radiation induces cellular deactivation: radiation tracks interacting with DNA can induce severe damage on its own ( $\alpha$  component) or can combine with another track to increase density of damage ( $\beta$  component).

Questions have, however, arisen to the relevance of the LQ model for more modern treatment regimens, such as SBRT or charged particle therapy. It has been shown that the LQ model begins to deviate significantly from experimental data beginning at or around 6–8 GyE (22, 90, 91). Practically speaking, this does not affect conventional treatment regimens utilizing 2–3 Gy fraction sizes; however, the LQ model may predict effects of hypofractionation regimes poorly. One such example is the carbon ion lung trial in Japan whereby single fractions of 50 GyE were delivered (92). Furthermore, when considering cases *in vivo*, the standard LQ model does not take into account repopulation kinetics of cancer cells during intra-fraction periods, rendering it approximate at best (93).

### **Modified LQ Models**

Modified versions of the canonical LQ model have been proposed to address some of the aforementioned shortcomings. Examples include generalized versions that have been further parameterized to account for repopulation (94), mixed radiation qualities (95), tumor heterogeneity, arbitrary or variable dose-rates (91), cell death mechanisms (96), and others able to take into account more than one of the aforementioned parameterizations (97).

In the case of charged particle therapy, the theory of dual radiation action (TDRA) predicts an increased linear component of LQ-modeled cell kill ( $\alpha$ ) over the quadratic component ( $\beta$ ) (98). Indeed, it has been shown that  $\beta$  remains relatively stable in comparison to the variation of  $\alpha$  across linear energy transfers (LETs) (99). Consequently, by considering the relative biological effectiveness (RBE) between a given high LET radiations and clinical energy photons, TDRA predicts that RBE will reach a minimum (RBE<sub>min</sub>) at very high doses, while at low doses RBE will reach an intrinsic maximum (RBE<sub>max</sub>) (100). In practice, it has been observed that both parameters  $\alpha$  and  $\beta$  vary with LET. Thus, a modified LQ model has emerged that RBE<sub>min</sub> and RBE<sub>max</sub>

are taken into account by further parameterizing the high-LET  $\alpha$  and  $\beta$  values (Eq. 4a, b).

$$\alpha_H = \alpha_L \cdot \text{RBE}_{\max} \quad (4a)$$

$$\beta_H = \beta_L \cdot \text{RBE}_{\min}^2 \quad (4b)$$

where  $\alpha_H$  and  $\alpha_L$  refer to  $\alpha$  components of the high and low LET radiations, respectively, and  $\beta_H$  and  $\beta_L$  refer to the quadratic components of the high and low LET radiations, respectively.

In considering high-dose per fraction, such as those delivered by a SBRT prostate radiotherapy plan or charged particle therapy, the modified LQ model first proposed by Sachs et al. in 1997 (101) and extended in 2004 by Guerrero and Li (91) has been demonstrated to fit survival data well (Eq. 5a, b).

$$SF = e^{-\alpha D - \beta \cdot G(\lambda T) \cdot D^2} \quad (5a)$$

$$G(\lambda T) = 2(\lambda T + e^{-\lambda T} - 1)/(\lambda T)^2 \quad (5b)$$

where  $\lambda$  is the repair rate and T is treatment delivery time while the other parameters are taken from the standard LQ model.  $G(\lambda T)$  is the dose protraction factor that specifies the contribution of misrepair to lethality (which is reduced at high, acute doses). Using this formulation of the LQ model, the large differences in cellular survival observed between predicted and experimental data are practically eliminated up to doses of  $\sim 30$  GyE (91).

Review articles by Jones and Dale (23) and Zaider and Hanin (102) have elegantly summarized and discussed additional models and can be consulted for further details.

### Normal Tissue Complication Probability

The objective of NTCP models is to gage the risk of inducing particular normal tissue effects, such as severe RB in the case of late prostate radiotherapy. A plethora of modeling techniques have been proposed for such purposes (103–105), some of which are described in more detail below.

#### Lyman–Kutcher–Burman

The most readily applied analytical method for generating NTCP values is through the Lyman–Kutcher–Burman (LKB) approach (106) (Eq. 6a,b).

$$\text{NTCP}(D, D_{50}, m) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{u^2}{2}} du \quad (6a)$$

$$\text{where } t = \frac{\text{EUD} - \text{TD}_{50}}{m \cdot \text{TD}_{50}} \text{ and } \text{TD}_{50}(V) = \frac{\text{TD}_{50}(1)}{V^n} \quad (6b)$$

where  $m$  is the slope of the best-fit NTCP sigmoid,  $\text{TD}_{50}(1)$  is the dose at which NTCP = 50% for a specific endpoint, and  $\text{TD}_{50}(V)$  is the tolerance dose for a given partial volume with tissue-specific volume exponent  $n$ . Simply put, the LKB model stratifies patient risk according to how much larger or smaller their EUD is relative to the  $\text{TD}_{50}$ . The EUD is a three-dimensional DVH reduction technique according to Eq. 7.

$$\text{EUD} = \left( \sum_i v_i \cdot D_i^{\frac{1}{a}} \right)^a \quad (7)$$

where parameter  $a$  is  $1/n$ , and  $D_i$  and  $v_i$  are the dose and partial volume, respectively, according to each DVH segment  $i$ . Expansions and more intricate variations of the canonical LKB model can be found readily throughout literature (57, 60, 62).

### Binomial Models

*The Critical Volume* Functional subunits (FSUs) are thought to be fundamentally underlying structured subunits housing numerous cells in a given tissue. Perhaps the most readily available example is the crypt subunit in the GI tract that together forms the organ. FSUs have varying properties, shapes, and sizes and are tissue type specific. Such variation can be exploited by the critical volume (CV) model (Eq. 8) to account for the differences in radiation response between different tissue types (107).

$$P_t = \binom{N}{t} P_{\text{FSU}}^t (1 - P_{\text{FSU}})^{N-t} \quad (8)$$

where the first term is the binomial coefficient of  $N$  and  $t$ .  $P_{\text{FSU}}$  is the probability that  $t$  of  $N$  subunits will be deactivated by ionizing radiation. Accordingly, the chance of  $M$  or more subunits being deactivated in a single exposure can be calculated according to:

$$P = \sum_{t=M+1}^N P_t = \sum_{t=M+1}^N \binom{N}{t} P_{\text{FSU}}^t (1 - P_{\text{FSU}})^{N-t}. \quad (9)$$

Two major classes of tissue exist in the context of FSUs (**Figures 1A,B**): serial and parallel. Organs that are serial can have their function compromised by exposure of a limited volume (a “critical volume”) to a given dose, e.g., colon, spinal cord, brain stem. Note that the output of a serial organ is not a sum of its internal components as it is in the case of parallel organs. Thus, for parallel-type tissues, catastrophic damage to one part of the underlying physiological architecture does not risk the collapse of the organ itself, e.g., liver, skin. In reality, every tissue has a mix of both serial and parallel structures, a concept referred to “complex” FSU arrangement (**Figure 1C**), although some tissues are more one type than the other.

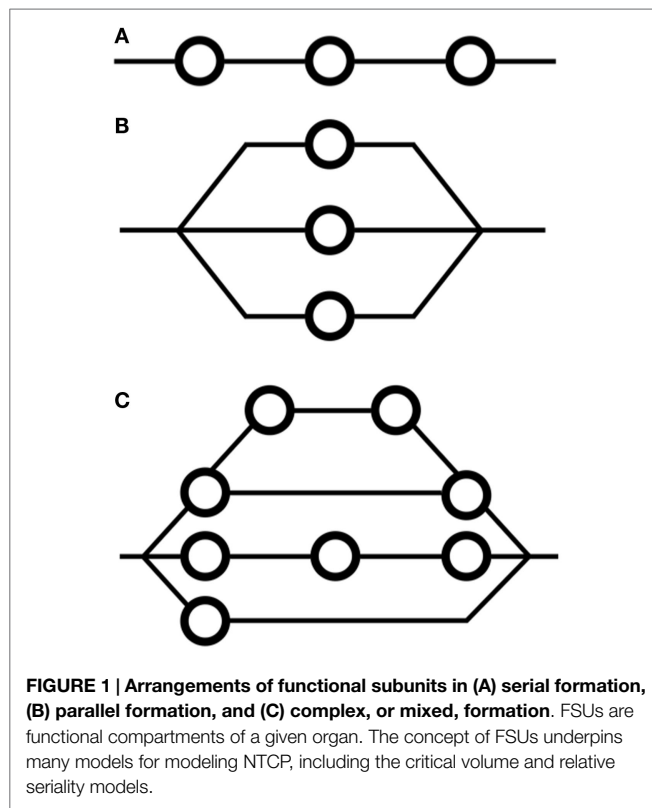
Interestingly, it has been shown that the LKB model can be derived upon reformulation of the CV model. This implies that the LKB model has a basis relating to FSUs (108).

Prior to mainstream applications of data-driven modeling techniques, modifications to the CV model (109, 110) and extensions of the FSU concept (111) were shown to be useful. Applications of the CV model in prostate cancer to predict late endpoints relating to bladder, colon, bowel, penile bulb, and rectum have been performed; however, their usefulness in practice has been limited compared to models identified using contemporary data-driven modeling approaches (112, 113).

*Relative Seriality* The relative seriality model for NTCP modeling was developed in order to consider and exploit arbitrary combinations of serial and parallel FSU arrangements (114). In such cases, risk of normal tissue damage is given by the following equation:

$$\text{NTCP}(D, V) = \left[ 1 - \left( 1 - P(D)^S \right)^{V/V_{\text{ref}}} \right]^{1/s} \quad (10)$$

where the exponent  $V/V_{\text{ref}}$  is the fraction of volume that is being irradiated to the given dose,  $D$ , and the parameter  $S$  relates to the degree of seriality of the organ at risk – nearly 0 in the case of highly parallel structures and higher for mixed or serial structures.



The function  $P(D)$  is the value of risk and can be derived via Poisson statistics:

$$P(D) = e^{-N_o \cdot S(D)} \quad (11)$$

where  $N_o$  is the number of FSUs and the function  $S(D)$  is the probability of a given FSU in the order of interest to survive irradiation to dose  $D$ .

## Data-Driven Modeling

Data-driven approaches to modeling are often referred to as phenomenological or statistical techniques. Models generated by such frameworks are based on empirical combinations of observations and are, thus, generally more robust than their analytical counterparts.

When considering dose–volume metrics alongside clinical and genetic risk factors, the number of variables can quickly become overwhelming and so data-driven modeling frameworks often include steps that seek to optimize and pre-process input data.

The most frequently employed approaches to data-driven modeling in radiotherapy are regression-based techniques. Regression link functions are typically chosen to be sigmoidal in order to achieve the non-linear dose–responses seen experimentally. Advanced methods in artificial intelligence that are able to handle non-linear data more readily are discussed later in this section. Such methods are becoming increasingly popular due to superior prospective classification performance in many areas of oncology (15, 115, 116).

Several review articles discussing the shortcomings and advantages of data-driven modeling can be found elsewhere in the literature (11, 117, 118).

## Probit- and Logit-Based Regression

Link functions can be used in tandem with regression frameworks to fit either TCP or NTCP data. The use of an inverse-logit (Eq. 12) or -probit (Eq. 13) are examples of such functions.

$$\pi(X_i) = \Phi(g(X_i)) \quad (12)$$

$$\pi(X_i) = \frac{e^{g(X_i)}}{1 + e^{g(X_i)}} = \frac{1}{1 + e^{-g(X_i)}} \quad (13)$$

where  $g(x_i)$  is the generalized linear model (GLM) formulation of the input variables:  $x_i$ :

$$g(X_i) = \beta_0 + \sum_{j=1}^s \beta_j X_{ij}, i = 1, \dots, n, j = 1, \dots, s \quad (14)$$

where  $\beta$  coefficients are estimated according to maximum likelihood estimation (MLE).

Historically, the logit function has been employed more often than the probit function because of ease of use and mathematical simplicity.

In terms of interpretation in the context of prostate radiotherapy, data-driven models have the added benefit of being able to handle multiple types of data while independently stratifying the contribution of specific variables. This can again be contrasted with analytical models wherein parameters need only be estimated rather than having to entirely develop the model itself.

## Artificial Intelligence (Machine Learning)

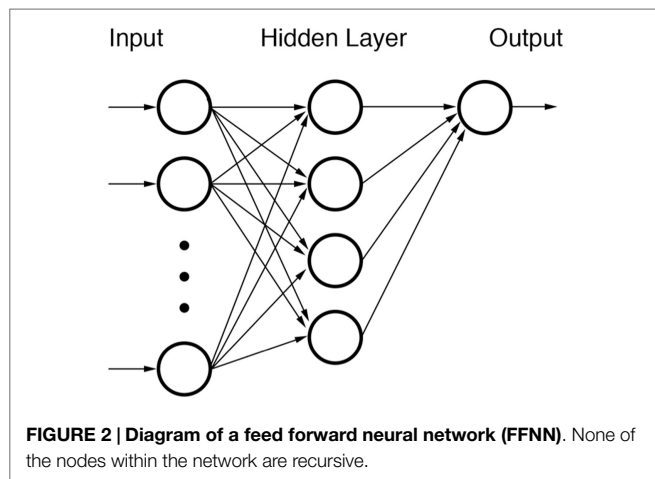
Techniques in artificial intelligence applied to outcome modeling consist of time-invariant statistical methods that are able, to a degree, to mimic selected human hallmarks. Artificial intelligence frameworks must first be able to learn (training phase) a pattern and then produce models that are able to recognize the pattern in a prospective setting (testing phase).

Success using artificial neural networks (ANNs), one of the major classes of artificial intelligence, has been achieved in learning and reproducing critical elements from the fields of speech pathology and handwriting recognition, both of which require complex recognition. Each node on a neural net indicates a function and, as such, refers to a transformation. In this context, a neural network itself carries no values without input data. In oncology, neural networks have also been used successfully although they have yet to be used prospectively (119–121).

Much criticism has arisen in recent years on the application of ANNs to prediction problems in oncology (122). Single hidden layer ANNs are universal function approximators, meaning that they can theoretically represent any function, which is defined by their topology and weighting values. This may lead to the fitting of implausible functions to datasets yielding uninterpretable and simply illogical results.

### Feed Forward Neural Network

Feed Forward Neural Networks (FFNNs) (Figure 2) do not include any recurring nodal inputs (“memories”) and are used frequently in basic pattern recognition problems. FFNNs are fully defined by their architecture such that arrangements of nodes into different topologies can induce different system responses. The



user decides what topology to employ for a given FFNN during the training phase although some have demonstrated the feasibility of using separate optimization algorithms to optimize network architecture itself (123–125). In radiation oncology, attempts have previously been made to utilize FFNNs for their ability to classify highly non-linear data (126, 127).

Each node of a FFNN represents a function with one or more inputs. Inputs from previous nodes are transformed according to an activation function. Examples of commonly used activation functions are logit or probit functions. Other functions, such as the radial basis function (RBF), can also be used if data are suspected of originating from a specific type of distribution. After transformation by activator functions, outputs from nodes are stratified by weights. Such weights are the elements of the FFNN that are trained when building a FFNN. Training of nodal weights is relatively straightforward albeit time-consuming. The *delta rule* can be used via back-propagation to adjust node input and output weights until classification performance is optimized. Datasets for training can be used all at once (batch training) or can be segregated into pattern-based subgroups (sequential training).

One shortcoming of FFNNs is that extraction of relationship data from within them can be notoriously difficult, if not, impossible. Information for final output nodes is reliant on previous inputs and outputs and, therefore, can become extremely mathematically complicated. This disadvantage is somewhat of a trade-off given that the only time-intensive procedure is training of node weights, after which the network can be used for real-time classification. Validated FFNNs are, therefore, indeed amenable to clinical implementation.

At least one group has demonstrated the applicability of ANNs in predicting late RB and in fact demonstrated improved prediction performance over and above that of regression-based approaches (123). Their findings exploited a genetic algorithm for optimizing inputs into their neural network and, furthermore, leveraged multiple cross-validation phases.

### Generalized Regression Neural Network

In contrast to the FFNN, the generalized regression neural network (GRNN) is a probabilistic neural network developed in 1991 and can, overall, be thought of as a best-fit estimator (128). The technique generalizes canonical regression by not being limited to

a specific function (e.g., in linear regression) but instead expresses an empirical regression function as a probability density determined using a technique known as *Parzen window estimation*. To accomplish this, the technique utilizes the joint probability of the input vector(s) and the outcomes to calculate conditional probabilities and expected values. These values are used to estimate the generalized regression of the outcomes onto the input data. The joint probability of the input vectors and outcomes can be estimated via non-parametric estimators if not known outright.

In the context of training, the advantage of GRNNs is that they avoid having to backpropagate error to fine tune nodal weights, which is computationally expensive and time-consuming. Back-propagation is mitigated by dealing with probability distributions rather than discrete raw input data. This means that one-pass of the framework with training data is sufficient to estimate parameter weights. Previous work by our group has shown that GRNNs can outperform FFNNs when it comes to prospective applications in radiation oncology (129). This is likely a result of the probabilistic nature of nature and/or biological variables across a patient cohort.

### Kernel-Based Methods

Kernel-based approaches to classification problems are based on clustering data according to non-linear combinations of variables (such as hyperplanes) in order to separate data. Oncology data are oftentimes highly non-linear, which gives motivation to explore the application of such a technique. Kernel methods seek to maximize distances between clusters that have undergone non-linear transformations. In this sense, the technique is a non-linear analog of Fischer's linear discriminant (FLD) analysis and principle component analysis (PCA) (Figure 3).

The most prominent member of the kernel-based learning family is the support vector machine (SVM). SVMs utilize support vectors that are formulated according to the most difficult to separate data and, therefore, are relied upon by the method in order to select an optimum classifier. By formulating the distance maximization problem between support vectors as a quadratic programming problem, a computationally efficient SVM formulation can be described by the following prediction function:

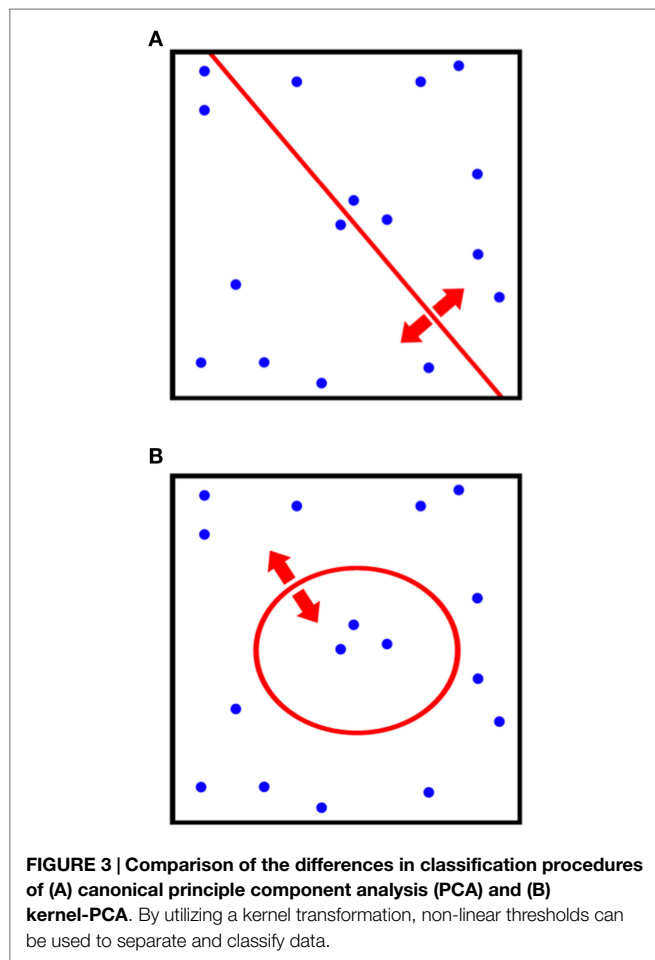
$$f(X) = \sum_{i=1}^{n_s} \alpha_i y_i K(S_i, X) + \alpha_o \quad (15)$$

where the number of support vectors is given by  $n_s$ ,  $K$  is the kernel transformation, and  $\alpha_i$  is the coefficient determined by quadratic programming. Further details on kernel-based methods can be found readily in the literature (130–132) or in our previous work (133).

### Systems Biology Approach

The concept of an integrated systems approach is that of understanding a given problem in terms of all of its components together, i.e., taking a “system-wide” view. This can be contrasted with a reductionist approach whereby each component of a system is looked at separately. A given system can be thought of having four principal components: structure (network topology), dynamics (time evolution of system), control (response and regulatory

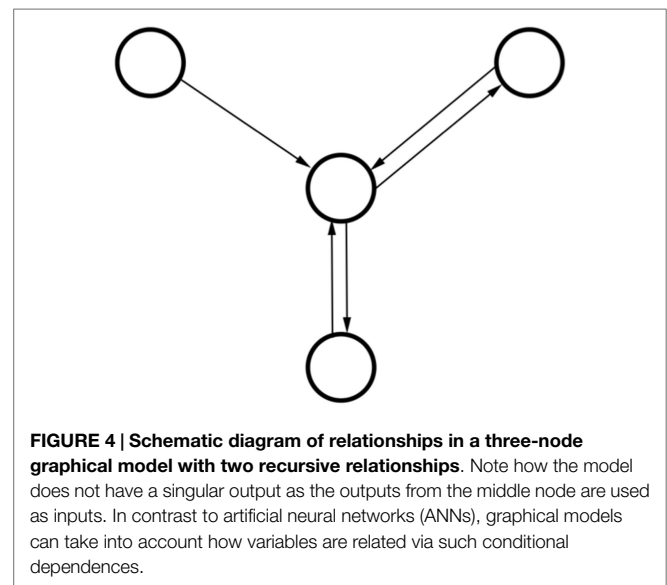




systems), and design (operational parameters or rules) (134). In applying a systems approach to biological and physiological systems, much insight could be gained that would otherwise be extremely challenging to extract using phenomenological models (135–137).

The biological effects induced by ionizing radiation are initiated at the atomic level in the form of free radical reactions. Free radical interactions occur rapidly and induce cascades of molecular responses, such as inflammation, which ultimately lead to recruitment of a variety of different molecular factors (138). Ultimately, over time, cellular responses can manifest themselves as clinical effects that are recorded as treatment outcomes, including toxicities. Much mechanistic insight and predictive power would be derived from a model that is able to combine these different organizational levels and related biophysical properties. Unfortunately, however, such a model requires expansive radiobiological knowledge spanning very different time and length scales, thus making the problem inherently complex.

In a related category, graphical models have been shown to be of use in radiation oncology as they can capture complex relationships between relevant factors and inter-dependencies between variables (Figure 4) (139, 140). Graphical models differ from aforementioned neural nets in that each random variable is represented by a node within the system and forms part of



an intricately connected web. The web simulates conditional probability relationships making this class of algorithm classified as a structured prediction technique as opposed to clustering or regression (discussed previously). In previous work, our group has shown good classification performance using a graphical Bayesian network in predicting radiation-induced pneumonitis (116).

## MODEL ORDER ESTIMATION

### Resampling Techniques

Frameworks that exploit resampling can be used to make estimates of model orders, parameters, or errors. In all cases, resampling requires that a dataset is repeatedly sampled with replacement in order to form many smaller, derived datasets. After several iterations, testing models on the derived datasets provides estimates on parameters of interest without requiring knowledge of the underlying distribution, which is a major benefit when little is known about the mechanics of the variable(s) of interest. Below, we discuss the two most commonly employed resampling techniques: jackknifing and bootstrapping.

#### Jackknifing

The jackknifing approach to parameter estimation entails systematically leaving out each of  $N$  samples and training  $N$  models on each of the  $N - 1$  remaining data points. The set of models trained on  $N - 1$  data points are then tested on the singular left-out data point one at a time. Analysis of the resulting  $N$  testing scores offers insight into how robust the model is under conditions of singular missing or inaccurate data (141, 142). Jackknifing is an approximation to the more labor-intensive, though robust, bootstrap technique.

#### Bootstrapping

“Bootstraps” are created from a given dataset by randomly resampling a given dataset (with replacement). Estimation of parameters or errors from each of the respective subsamples (a process known as bagging) can then be performed and averaged to

yield an average value. The method is simple and not limited to specific classes of parameters and is, therefore, an extremely flexible technique (116, 118).

Bootstrapping is often used in cases where analytical error estimation is unfeasible. One shortcoming of bootstrap resampling is that it assumes independence of data points. In the context of radiotherapy outcomes, each set of data points originates from a specific patient and so this is often not an issue.

## Information Theory Approaches

Theoretical approaches to order estimation based on concepts borrowed from information theory can be used as alternatives to resampling techniques. These methods serve as tools to help identify which out of a finite number of models explain outcome data best. Therefore, the two methods discussed herein only indicate relative measures of fit and do not give any information on the quality of any model in an absolute sense (143).

### Akaike Information Criteria

The Akaike information criteria (AIC) is based on the principle of good-of-fit and penalizes models that over- or under-fit data (144). The approach utilizes the Kullback–Leibler (K–L) distance, which quantifies the difference between two probability distributions. The K–L distance is used to estimate divergence of potential models from their true sampling distributions. The AIC approach furthermore rewards the models with the lowest value of the AIC parameter, which is calculated by considering the likelihood ( $L$ ) of a particular model to explain outcome data (Eq. 16).

$$AIC = 2k - \ln(L) \quad (16)$$

where  $k$  is the number of parameters in the model. In order to find the optimal AIC for a given set of models, the equation is minimized via maximizing the log-likelihood term on the right-hand side. The additional term  $2k$  is a penalty factor that penalizes over-fitting of data with increased number of variables. One shortcoming of AIC is that it can fail when large numbers of models are under examination due to the multiple comparison dilemma.

### Bayesian Information Criteria

The Bayesian information criteria (BIC) is a closely related concept to AIC and can similarly provide information as to which out of many models best explains a given set of data best (145). The BIC is based on Bayesian inference and is formally given by:

$$BIC = k \cdot \ln(n) - 2 \cdot \ln(L) \quad (17)$$

where  $L$  is the maximum of the likelihood function,  $k$  is the number of parameters in the model, and  $n$  is the number of data points i.e., sample size. Threshold values for BIC that decidedly indicate whether a particular model should be discarded have been composed and tested by Kass and Raftery (p. 777) (146).

In comparison to the AIC, the BIC has a larger penalty term  $k \ln(n)$  and, thus, penalize over-fitting more than does the AIC. As a result, BIC prefers models with fewer parameters than those chosen by AIC. The BIC also suffers when  $k$  is large due to the high-dimensionality problem of identifying variables that fit by chance.

## EVALUATION OF MODEL PERFORMANCE

There exist numerous methods in literature to evaluate the ability of a given model to classify data in a prospective sense. Oftentimes, frameworks will employ more than one validation technique in order to explore the shortcomings of outputted models.

### Validation Coefficients and Metrics

Metrics and coefficients are the most readily available tools for calculating the prediction or classification performance of outcome models. Their simplicity is amenable to quick understanding of model behavior and, when several are used together, can yield insightful information.

The linear Pearson's correlation is an example of a non-parametric coefficient that is used frequently for estimating the linearity of a relationship between two variables. More often employed in outcome models is the Spearman Rank Coefficient, which does not assume linearity and instead yields an estimate on the direction of trend between two parameters.

Alternatively, receiver-operating characteristic (ROC) values can be summed from ROC plots to readily convey classification performance alongside sensitivity and specificity for the desired classification cutoff value.

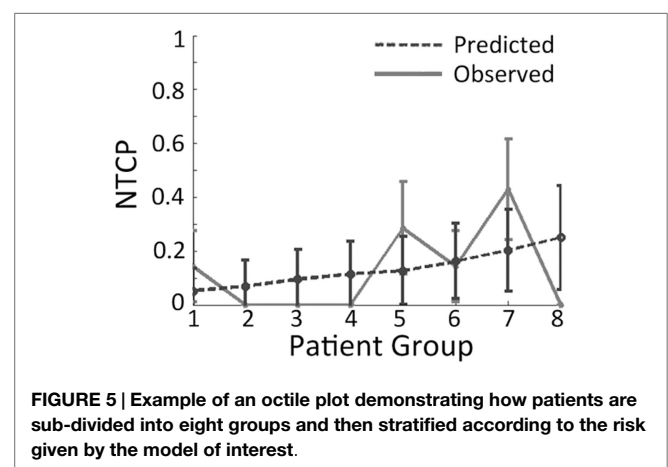
### Cross-Validation by Resampling

Resampling with replacement can be used to quantify the classification performance of models as well as estimate confidence intervals on model performance or provide estimate on the error of classification statistics. In our experience, leave-one-out cross-validation (LOOCV) on finalized models serves as an excellent method to quickly estimate how robust a given model is without having to rely on more computationally expensive methods, such as bootstrapping.

## MODEL PERFORMANCE VISUALIZATION

### Octile Plots

Plots whereby outcomes are split into eight groups (octiles) are called octile plots (Figure 5). By considering and plotting both the predicted and observed outcomes, the plot provides a two-dimensional method to visually assess model fit. Furthermore,



**FIGURE 5 |** Example of an octile plot demonstrating how patients are sub-divided into eight groups and then stratified according to the risk given by the model of interest.

octile plots allow the reader to gage how the overall model output varies with increasing magnitude of input parameters.

## Receiver-Operating Characteristic Plot

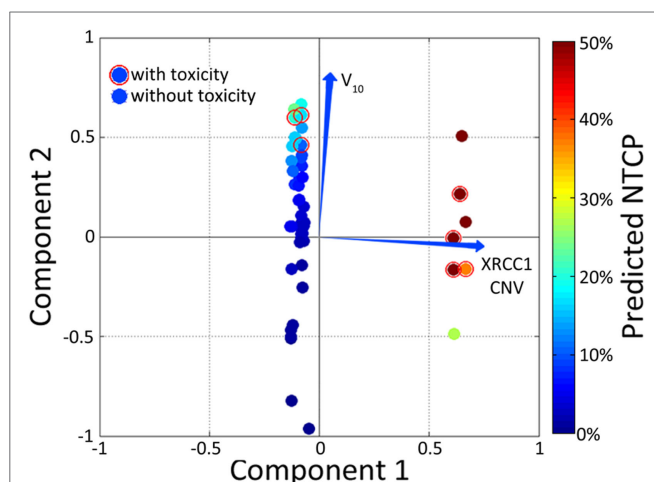
As discussed previously, the ROC plot can provide a method for visually assessing model performance. Although the AUC parameter derived from the ROC plot is reported more often, the ROC plot itself is also useful as it provides information on how the sensitivity varies with the specificity for different threshold cut-offs. In our work, we often use ROC plots alongside AUCs and correlation coefficients to gain a full understanding of how a particular model behaves under conditions of cross-validation (60).

## Vector Biplots

Biplots display vectors that are constructed and presented alongside PCA-derived information and patient-specific NTCP. In the context of model visualization, vector biplots provide a rough estimate of which variables are able to explain the data – this is generally accomplished in two dimensions to aid visualizing the model on paper.

## PCA and Kernel-PCA

Principle component analysis is useful as a method to reduce the dimensionality of data into either two- or three-dimensions to facilitate performance visualization. The trade-off in using PCA for outcome modeling is that the relationships between inputs and outcomes are highly non-linear and, therefore, true relationships may not be adequately captured by the technique. Alternatively, the previously described Kernel-PCA technique can be employed to visualize data by improving separation between clusters. Vector biplots, two- and three-dimensional kPCA plots can indeed be used together in order to provide easy-to-interpret heat maps colored by estimated patient-specific risk (Figure 6) (147).



**FIGURE 6 | Example of a color-washed vector biplot.** The cardinal plane represents the vector biplot relating to the magnitude of the contribution of the variables contained within the model. The axes on which the vector biplot is shown are derived via principle component analysis (PCA). Dummy patients with toxicity were circled with empty red circles and color-washed according to NTCP values generated via the model.

## CONTROVERSIES

The number of variables in a big data analysis of outcome models can accumulate quickly, especially if considering many biological factors. Several of the modeling techniques presented herein fail to take into account or mitigate the dilemma of multiple comparisons of chance correlations. Therefore, in all aspects of modeling, independent validation phases should be integrated into frameworks that aspire to produce clinically relevant results. Furthermore, although internal cross-validation techniques do provide excellent estimates of model robustness, their usefulness is limited if training samples are not retrieved from independent sources.

Radiotherapy outcomes are complex pathological manifestations of the biophysical effects of ionizing radiation on the human body. Therefore, models attempting to delineate such phenomena should endeavor to incorporate as many different types as possible.

By definition, big data requires the ability to work with extremely large multi-dimensional datasets, which requires dedicated infrastructure to support access as well as high performance capabilities to facilitate efficient data exploration and modeling. Such investments require capital expenditure and training in addition to the formation of data-sharing and privacy agreements between institutions (148).

Once meaningful analytics can readily be extracted from multicenter databases, scientists and physicians are faced with the dilemma of determining how their predictive models should be used to maximize the TCP/NTCP ratio. No doubt, the aforementioned biological and clinical risk factors that predispose a prostate cancer patient to radiation toxicity can be quantified before therapy and used to guide initial radiation treatment planning but such an approach ignores such factors as risk to long-term quality of life, intra-treatment imaging data, physiological changes during therapy, and symptoms the patient may develop during and shortly after therapy.

## SOFTWARE TOOLS

Many independently verified platforms exist for data-mining and analytics exploration, several of which are listed below with brief explanations of their scope and limitations. Further details on the use and QA of such software can be found in AAPM Task Group Report #166 (149).

### BIOlogical Evaluation of PLANs (BIOPLAN)

Bioplan is a user-friendly software developed in the United Kingdom that allows an absorbed dose treatment plan to be converted into its likely biological effect (150). It provides a variety of tools, including DVH subtraction, and is able to calculate NTCP values according to previously described LKB and binomial-based models.

### Computational Environment for Radiotherapy Research

Computational environment for radiotherapy research (CERR) is an open-source computational environment that facilitates the

conversion of treatment plan data into MATLAB (151). The software allows for either retrospective or experimental treatment planning and can read-in CT data as well as associated contours. In the past, our group has used CERR in investigating both GU and GI toxicities in prostate cancer. CERR can also be used to estimate the contribution of joint-contours, such as rectal margin overlap with the PTV, to toxicities.

## Dose Response Explorer System

Dose response explorer system (DREES) is an open-source data-mining tool for exploring dose–response relationships (142). Using a built-in subroutine called CERR+, the program imports patient data from CERR. DREES provides a suite of tools for either NTCP or TCP modeling of outcomes without restriction as to the site or population size. The program includes a GUI interface within MATLAB in order to simplify usability. Examples of the functions contained with DREES include logistic regression, LKB modeling, actuarial statistics, bootstrap validation, Kaplan–Meier survival analysis, nomograms, boxplots, and has the functionality for interfacing with SVMs. One of the major advantages of using DREES is it that it freely distributed, as is CERR, and is consistently updated, making it flexible and adaptable.

## FUTURE TRENDS

### Implications of Charged Particle Therapy

The use of charged particle therapy in treating prostate cancer has become more mainstream over the past decade, mainly in the United States (152). Unfortunately, comparable outcome data from a treatment delivery technology point of view on the use of proton therapy vs. photon therapy for prostate cancer (to an even lesser extent carbon ion therapy) are not yet readily available. Unlike the mainstream adoption of RBE = 1.1 for proton therapy (153), the RBE debate for charged particles continues, which makes outcomes comparison to photons difficult for heavier charged particles and prospective outcomes prediction impossible.

Studies published using modulated proton techniques report that GI/GU late effects post-RT are either unchanged (154) or reduced (155–157). In the case of carbon ions, extensive long-term biochemical outcomes are not yet available and so does the interpretation of the efficacy of such treatments remains difficult. In terms of late radiation-induced ED, at least one study has shown no significant upregulation using protons (158).

The impact of charged particle therapy plans on outcome modeling is that thresholds, such as those proposed by QUANTEC, will likely need to be adjusted given the differences in dose distribution and biological response relative to photon-based therapy (159, 160).

### Advanced Methods in Machine Learning

Examples of more complex modeling techniques include the use of restricted Boltzmann machines (RBMs), which are energy-based multilayered graphical models that estimate the joint probability distribution between inputs and outcomes using one or more binary stochastic hidden layers (161–163). The concept of

an RBM was first proposed by Prof. Geoffrey Hinton from the University of Toronto in 2006 as a method to efficiently train and learn a constrained version of a Boltzmann machine (164). RBMs as applied to oncology have shown promise in making accurate predictions, however, they remain relatively poorly disseminated techniques and their implementation to date has been limited.

Multilayered networks, such as RBM or convolutional neural network, can perform *deep learning* in that data are passed through more than one layers of machine-learning modules that combine to form a framework capable of processing highly complex patterns. More specifically, deep learning strategies attempt to model high-level abstractions, such as the recognition of three-dimensional objects, in order to label and classify. This may prove extremely useful in the case of radiation-induced biological effects given the physical understanding that such phenomena manifest after cascading effects at the atomic, molecular, and then cellular levels. Previously, modeling strategies involving deep learning in the form of deep belief networks (DBNs) have been applied to data in oncology via multilayered RBMs, which are known as deep Boltzmann machines (DBMs) (165, 166). Such applications are, for instance, infrequent but may become more frequent as techniques are disseminated and refined specifically for the purposes of radiation oncology.

## CONCLUSION

Prostate cancer is one of the most commonly diagnosed cancers across the globe and radiation therapy is a primary modality in treating such cases. Prostate cancer is a heterogeneous disease at several levels (pathological, molecular/genetic, and clinical) and, despite technical improvements, there is still a significant risk of cancer recurrence after therapy. Treatment efficacy for localized prostate cancer has increased greatly in recent years; however, few efforts have been aimed at developing and testing personalized predictive metrics, namely those identified through the rapidly advancing field of big data analytics using machine learning and artificial intelligence. Such analytics would allow prostate radiotherapy regimens to be further tailored to the individual and generate treatment plans that are functions of not only dose–volume metrics but also individual's clinical risk factors and biological parameters.

Given the apparent complexity of physiological response to ionizing radiation, it is likely that systems-based approaches will play a larger role in radiotherapy outcome modeling in the future. Although regression-based techniques have yielded success in certain cases, their cross-validated prediction performance appears to be generally limited, likely due to their inability to capture higher order interactions between biophysical processes.

A chief limitation in modern outcome modeling projects is the difficulty in pooling sets of data from multiple institutions. In medicine, this is principally due to privacy and security concerns. If, however, proper data-sharing protocols can be put in place, big data analytics may provide a significant boost to data-driven outcome models owing not only to larger datasets but also to the ability to more readily perform independent validation.

## AUTHOR CONTRIBUTIONS

The initial conception of the review article was proposed by IN and the initial draft was compiled by JC. Multiple rounds of peer

editing by both JC and IN then took place until LS was contacted at the request of the editors in order to review the manuscript for clinical input. All figures were selected and compiled together by JC, LS, and IN.

## REFERENCES

- UK CR. Prostate cancer statistics. *Cancer Stat* (2011). Available from: <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer#heading-Zero>
- Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* (2013) 49:1374–403. doi:10.1016/j.ejca.2012.12.027
- Network NCI. *Treatment Routes in Prostate Cancer*. Bristol, UK (2009).
- Sriprasad S, Feneley MR, Thompson PM. History of prostate cancer treatment. *Surg Oncol* (2009) 18:185–91. doi:10.1016/j.suronc.2009.07.001
- Bauman G, Rumble RB, Chen J, Loblaw A, Warde P, Members of the IMRT Indications Expert Panel. Intensity-modulated radiotherapy in the treatment of prostate cancer. *Clin Oncol (R Coll Radiol)* (2012) 24:461–73. doi:10.1016/j.clon.2012.05.002
- Musunuru HB, Loblaw A. Clinical trials of stereotactic ablative radiotherapy for prostate cancer: updates and future direction. *Future Oncol* (2015) 11:819–31. doi:10.2217/fo.15.14
- De Ruyscher D, Mark Lodge M, Jones B, Brada M, Munro A, Jefferson T, et al. Charged particles in radiotherapy: a 5-year update of a systematic review. *Radiother Oncol* (2012) 103:5–7. doi:10.1016/j.radonc.2012.01.003
- Hummel S, Simpson EL, Hemingway P, Stevenson MD, Rees A. Intensity-modulated radiotherapy for the treatment of prostate cancer: a systematic review and economic evaluation. *Health Technol Assess* (2010) 14:1–108,iii–iv. doi:10.3310/hta14470
- Milano MT, Constine LS, Okunieff P. Normal tissue tolerance dose metrics for radiation therapy of major organs. *Semin Radiat Oncol* (2007) 17:131–40. doi:10.1016/j.semradonc.2006.11.009
- Mundt A, Boeske JC. *Intensity Modulated Radiation Therapy: A Clinical Perspective*. Tokyo (2005). p. 151–2.
- Marks LB, Yorke ED, Jackson A, Ten Haken RK, Constine LS, Eisbruch A, et al. Use of normal tissue complication probability models in the clinic. *Int J Radiat Oncol Biol Phys* (2010) 76:S10–9. doi:10.1016/j.ijrobp.2009.07.1754
- Bentzen SM, Parliament M, Deasy JO, Dicker A, Curran WJ, Williams JP, et al. Biomarkers and surrogate endpoints for normal-tissue effects of radiation therapy: the importance of dose-volume effects. *Int J Radiat Oncol Biol Phys* (2010) 76:S145–50. doi:10.1016/j.ijrobp.2009.08.076
- Deasy JO, Mayo CS, Orton CG. Treatment planning evaluation and optimization should be biologically and not dose/volume based. *Med Phys* (2015) 42:2753–6. doi:10.1118/1.4916670
- El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *Int J Radiat Oncol Biol Phys* (2006) 64:1275–86. doi:10.1016/j.ijrobp.2005.11.022
- Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48:441–6. doi:10.1016/j.ejca.2011.11.036
- Rosenstein BS, West CM, Bentzen SM, Alsner J, Andreassen CN, Azria D, et al. Radiogenomics: radiobiology enters the era of big data and team science. *Int J Radiat Oncol Biol Phys* (2014) 89:709–13. doi:10.1016/j.ijrobp.2014.03.009
- Keetch DW, Humphrey PA, Smith DS, Stahl D, Catalona WJ. Clinical and pathological features of hereditary prostate cancer. *J Urol* (1996) 155:1841–3. doi:10.1016/S0022-5347(01)66024-5
- Epstein JI, Walsh PC, Carmichael M, Brendler CB. Pathologic and clinical findings to predict tumor extent of nonpalpable (stage T1c) prostate cancer. *JAMA* (1994) 271:368–74. doi:10.1001/jama.271.5.368
- Heppner GH, Miller BE. Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metastasis Rev* (1983) 2:5–23. doi:10.1007/BF00046903
- Mackinnon AC, Yan BC, Joseph LJ, et al. Molecular biology underlying the clinical heterogeneity of prostate cancer: an update. *Arch Pathol Lab Med* (2009) 133:1033–40. doi:10.1043/1543-2165-133.7.1033
- Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* (2010) 18:11–22. doi:10.1016/j.ccr.2010.05.026
- Kirkpatrick JP, Meyer JJ, Marks LB. The linear-quadratic model is inappropriate to model high dose per fraction effects in radiosurgery. *Semin Radiat Oncol* (2008) 18:240–3. doi:10.1016/j.semradonc.2008.04.005
- Jones B, Dale RG. Mathematical models of tumour and normal tissue response. *Acta Oncol* (1999) 38:883–93. doi:10.1080/028418699432572
- Brenner DJ, Hall EJ. Fractionation and protraction for radiotherapy of prostate carcinoma. *Int J Radiat Oncol Biol Phys* (1999) 43:1095–101. doi:10.1016/S0360-3016(98)00438-6
- Fowler J, Chappell R, Ritter M. Is  $\alpha/\beta$  for prostate tumors really low? *Int J Radiat Oncol Biol Phys* (2001) 50:1021–31. doi:10.1016/S0360-3016(01)01607-8
- Fowler JF. The radiobiology of prostate cancer including new aspects of fractionated radiotherapy. *Acta Oncol* (2005) 44:265–76. doi:10.1080/02841860410002824
- Dasu A. Is the  $\alpha/\beta$  value for prostate tumours low enough to be safely used in clinical trials? *Clin Oncol* (2007) 19:289–301. doi:10.1016/j.clon.2007.02.007
- Kupelian PA, Thakkar VV, Khuntia D, Reddy CA, Klein EA, Mahadevan A. Hypofractionated intensity-modulated radiotherapy (70 Gy at 2.5 Gy per fraction) for localized prostate cancer: long-term outcomes. *Int J Radiat Oncol Biol Phys* (2005) 63:1463–8. doi:10.1016/j.ijrobp.2005.05.054
- Arcangeli G, Saracino B, Gomellini S, Petrongari MG, Arcangeli S, Sentinelli S, et al. A prospective phase III randomized trial of hypofractionation versus conventional fractionation in patients with high-risk prostate cancer. *Int J Radiat Oncol Biol Phys* (2010) 78:11–8. doi:10.1016/j.ijrobp.2009.07.1691
- King CR, Brooks JD, Gill H, Pawlicki T, Cotrutz C, Presti JC. Stereotactic body radiotherapy for localized prostate cancer: interim results of a prospective phase II clinical trial. *Int J Radiat Oncol Biol Phys* (2009) 73:1043–8. doi:10.1016/j.ijrobp.2008.05.059
- Vogelius IR, Bentzen SM. Meta-analysis of the alpha/beta ratio for prostate cancer in the presence of an overall time factor: bad news, good news, or no news? *Int J Radiat Oncol Biol Phys* (2013) 85:89–94. doi:10.1016/j.ijrobp.2012.03.004
- Dörr W, Hendry JH. Consequential late effects in normal tissues. *Radiother Oncol* (2001) 61:223–31. doi:10.1016/S0167-8140(01)00429-7
- Pinkawa M, Holy R, Piroth MD, Fishedick K, Schaar S, Székely-Orbán D, et al. Consequential late effects after radiotherapy for prostate cancer – a prospective longitudinal quality of life study. *Radiat Oncol* (2010) 5:27. doi:10.1186/1748-717X-5-27
- Pinkawa M, Ribbing C, Djukic V, Klotz J, Holy R, Eble MJ. Early hematologic changes during prostate cancer radiotherapy predictive for late urinary and bowel toxicity. *Strahlentherapie Und Onkol Organ Der Dtsch Röntgengesellschaft* (2015) 191:771–7. doi:10.1007/s00066-015-0841-3
- Droge CS, Boychak O, Patel S, Usmani N, Amanie J, Parliament MB, et al. Acute toxicity of hypofractionated intensity-modulated radiotherapy for prostate cancer. *Curr Oncol* (2015) 22:e76–84. doi:10.3747/co.22.2247
- Zietman AL. The Titanic and the iceberg: prostate proton therapy and health care economics. *J Clin Oncol* (2007) 25:3565–6. doi:10.1200/JCO.2007.11.9768
- Yu JB, Soulos PR, Herrin J, Cramer LD, Potosky AL, Roberts KB, et al. Proton versus intensity-modulated radiotherapy for prostate cancer: patterns of care and early toxicity. *J Natl Cancer Inst* (2013) 105(1):25–32. doi:10.1093/jnci/djs463
- Fang P, Mick R, Deville C, Both S, Bekelman JE, Christodouleas JP, et al. A case-matched study of toxicity outcomes after proton therapy and intensity-modulated radiation therapy for prostate cancer. *Cancer* (2015) 121:1118–27. doi:10.1002/cncr.29148

39. Coen JJ, Bae K, Zietman AL, Patel B, Shipley WU, Slater JD, et al. Acute and late toxicity after dose escalation to 82 GyE using conformal proton radiation for localized prostate cancer: initial report of American College of Radiology Phase II Study 03-12. *Int J Radiat Oncol Biol Phys* (2011) 81:1005–9. doi:10.1016/j.ijrobp.2010.06.047
40. Hoeller U, Tribius S, Kuhlmeier A, Grader K, Fehlaue F, Alberti W. Increasing the rate of late toxicity by changing the score? A comparison of RTOG/EORTC and LENT/SOMA scores. *Int J Radiat Oncol Biol Phys* (2003) 55:1013–8. doi:10.1016/S0360-3016(02)04202-5
41. Faria SL, Aslani M, Tafazolli FS, Souhami L, Freeman CR. The challenge of scoring radiation-induced lung toxicity. *Clin Oncol* (2009) 21:371–5. doi:10.1016/j.clon.2009.01.017
42. Denis F, Garaud P, Bardet E, Alfonsi M, Sire C, Germain T, et al. Late toxicity results of the GORTEC 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: comparison of LENT/SOMA, RTOG/EORTC, and NCI-CTC scoring systems. *Int J Radiat Oncol Biol Phys* (2003) 55:93–8. doi:10.1016/S0360-3016(02)03819-1
43. Bruce JY, Lang JM, McNeel DG, Liu G. Current controversies in the management of biochemical failure in prostate cancer. *Clin Adv Hematol Oncol* (2012) 10:716–22.
44. Consensus statement: guidelines for PSA following radiation therapy. American Society for Therapeutic Radiology and Oncology Consensus Panel. *Int J Radiat Oncol Biol Phys* (2015) 37:1035–41. doi:10.1016/S0360-3016(97)00002-3
45. Barker JL Jr, Garden AS, Ang KK, O'Daniel JC, Wang H, Court LE, et al. Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system. *Int J Radiat Oncol Biol Phys* (2004) 59:960–70. doi:10.1016/j.ijrobp.2003.12.024
46. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. *Med Phys* (1997) 24:103–10. doi:10.1118/1.598154
47. Niemierko A. A generalized concept of equivalent uniform dose (EUD). *Med Phys* (1999) 26:1100.
48. Hope AJ, Lindsay PE, El Naqa I, Alaly JR, Vicic M, Bradley JD, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int J Radiat Oncol Biol Phys* (2006) 65:112–24. doi:10.1016/j.ijrobp.2005.11.046
49. Acosta O, Dowling J, Cazoulat G, Simon A, Salvado O, de Crevoisier R, et al. Atlas based segmentation and mapping of organs at risk from planning CT for the development of voxel-wise predictive models of toxicity in prostate radiotherapy. In: Madabhushi A, Dowling J, Yan P, Fenster A, Abolmaesumi P, Hata N, editors. *Prostate Cancer Imaging. Comput. Diagnosis, Progn. Interv. SE – 6*. (Vol. 6367), Berlin: Springer (2010). p. 42–51.
50. Sivanathan S, Jang S, Mittal BB, Luo J, Grein CH, Klie R, et al. A study of the radiobiological modeling of the conformal radiation therapy in cancer treatment. (2012). Available at: <https://indigo.uic.edu/handle/10027/9467>
51. Hakimi AA, Ostrovskaya I, Reva B, Schultz N, Chen YB, Gonen M, et al. Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. *Clin Cancer Res* (2013) 19:3259–67. doi:10.1158/1078-0432.CCR-12-3886
52. Svensson JP, Stalpers LJA, Lange REEE, Franken NAP, Haveman J, Klein B, et al. Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity. *PLoS Med* (2006) 3:e422. doi:10.1371/journal.pmed.0030422
53. Barnett GC, Elliott RM, Alsner J, Andreassen CN, Abdelhay O, Burnet NG, et al. Individual patient data meta-analysis shows no association between the SNP rs1800469 in TGFB and late radiotherapy toxicity. *Radiother Oncol* (2012) 105:289–95. doi:10.1016/j.radonc.2012.10.017
54. Kerns SL, Ostrer H, Stock R, Li W, Moore J, Pearlman A, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys* (2010) 78:1292–300. doi:10.1016/j.ijrobp.2010.07.036
55. Burri RJ, Stock RG, Cesaretti JA, Atencio DP, Peters S, Peters CA, et al. Association of single nucleotide polymorphisms in SOD2, XRCC1 and XRCC3 with susceptibility for the development of adverse effects resulting from radiotherapy for prostate cancer. *Radiat Res* (2008) 170:49–59. doi:10.1667/RR1219.1
56. Barnett GC, Coles CE, Burnet NG, Pharoah PDP, Wilkinson J, West CML, et al. No association between SNPs regulating TGF- $\beta$ 1 secretion and late radiotherapy toxicity to the breast: results from the RAPPER study. *Radiother Oncol* (2010) 97:9–14. doi:10.1016/j.radonc.2009.12.006
57. Defraene G, Van Den Bergh L, Al-Mamgani A, Haustermans K, Heemsbergen W, Van Den Heuvel F, et al. The benefits of including clinical factors in rectal normal tissue complication probability modeling after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys* (2012) 82:1233–42. doi:10.1016/j.ijrobp.2011.03.056
58. Parliament MB, Murray D. Single nucleotide polymorphisms of DNA repair genes as predictors of radioresponse. *Semin Radiat Oncol* (2010) 20:232–40. doi:10.1016/j.semradi.2010.05.003
59. Barnett GC, West CML, Dunning AM, Elliott RM, Coles CE, Pharoah PDP, et al. Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype. *Nat Rev Cancer* (2009) 9:134–42. doi:10.1038/nrc2587
60. Coates J, Jeyaseelan AK, Ybarra N, David M, Faria S, Souhami L, et al. Contrasting analytical and data-driven frameworks for radiogenomic modeling of normal tissue toxicities in prostate cancer. *Radiother Oncol* (2015) 115(1):107–13. doi:10.1016/j.radonc.2015.03.005
61. Peeters JV, Hoogeman MS, Heemsbergen WD, Hart AAM, Koper PCM, Lebesque JH. Rectal bleeding, fecal incontinence, and high stool frequency after conformal radiotherapy for prostate cancer: normal tissue complication probability modeling. *Int J Radiat Oncol Biol Phys* (2006) 66:11–9. doi:10.1016/j.ijrobp.2006.03.034
62. Tucker SL, Li M, Xu T, Gomez D, Yuan X, Yu J, et al. Incorporating single-nucleotide polymorphisms into the Lyman model to improve prediction of radiation pneumonitis. *Int J Radiat Oncol Biol Phys* (2013) 85:251–7. doi:10.1016/j.ijrobp.2012.02.021
63. Rancati T, Fiorino C, Fellin G, Vavassori V, Cagna E, Casanova Borca V, et al. Inclusion of clinical risk factors into NTCP modelling of late rectal toxicity after high dose radiotherapy for prostate cancer. *Radiother Oncol* (2011) 100:124–30. doi:10.1016/j.radonc.2011.06.032
64. Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. *Nat Med* (2011) 17:330–9. doi:10.1038/nm.2305
65. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* (2009) 31:27–36. doi:10.1093/carcin/bgp220
66. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell* (2012) 150:12–27. doi:10.1016/j.cell.2012.06.013
67. Smits KM, Melotte V, Niessen HEC, Dubois L, Oberije C, Troost EGC, et al. Epigenetics in radiotherapy: where are we heading? *Radiother Oncol* (2014) 111:168–77. doi:10.1016/j.radonc.2014.05.001
68. Muscarella LA, Parrella P, D'Alessandro V, la Torre A, Barbano R, Fontana A, et al. Frequent epigenetics inactivation of KEAP1 gene in non-small cell lung cancer. *Epigenetics* (2011) 6:710–9. doi:10.4161/epi.6.6.15773
69. Li LC, Carroll PR, Dahiya R. Epigenetic changes in prostate cancer: implication for diagnosis and treatment. *J Natl Cancer Inst* (2005) 97:103–15. doi:10.1093/jnci/dji010
70. Li L-C. Epigenetics of prostate cancer. *Front Biosci* (2007) 12:3377–97. doi:10.2741/2320
71. Nightingale KP, O'Neill LP, Turner BM. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev* (2006) 16:125–36. doi:10.1016/j.gde.2006.02.015
72. Turner BM. Defining an epigenetic code. *Nat Cell Biol* (2007) 9:2–6. doi:10.1038/ncb0107-2
73. Turner BM. Histone acetylation and an epigenetic code. *Bioessays* (2000) 22:836–45. doi:10.1002/1521-1878(200009)22:9<836:AID-BIES9>3.0.CO;2-X
74. Rakan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* (2011) 12:529–41. doi:10.1038/nrg3000
75. Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology – multifactorial decision support systems. *Nat Rev Clin Oncol* (2013) 10:27–40. doi:10.1038/nrclinonc.2012.196
76. Abd El-Rehim DM, Ball G, Finder SE, Rakha E, Paish C, Robertson JFR, et al. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int J Cancer* (2005) 116:340–50. doi:10.1002/ijc.21004

77. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* (2011) 29:17–24. doi:10.1200/JCO.2010.30.1077
78. Bremnes RM, Veve R, Gabrielson E, Hirsch FR, Baron A, Bemis L, et al. High-throughput tissue microarray analysis used to evaluate biology and prognostic significance of the E-cadherin pathway in non-small-cell lung cancer. *J Clin Oncol* (2002) 20:2417–28. doi:10.1200/JCO.2002.08.159
79. Kulshreshtha R, Ferracin M, Wojcik SE, Garzon R, Alder H, Agosto-Perez FJ, et al. A microRNA signature of hypoxia. *Mol Cell Biol* (2007) 27:1859–67. doi:10.1128/MCB.01395-06
80. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br J Cancer* (2010) 102:428–35. doi:10.1038/sj.bjc.6605450
81. Chi J-T, Wang Z, Nuyten DSA, Rodriguez EH, Schaner ME, Salim A, et al. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med* (2006) 3:e47. doi:10.1371/journal.pmed.0030047
82. Clarke R, Ransom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* (2008) 8:37–49. doi:10.1038/nrc2294
83. Alaiya AA, Franzén B, Auer G, Linder S. Cancer proteomics: from identification of novel markers to creation of artificial learning models for tumor classification. *Electrophoresis* (2000) 21:1210–7. doi:10.1002/(SICI)1522-2683(20000401)21:6<1210::AID-ELPS1210>3.0.CO;2-S
84. Kreso A, O'Brien CA, van Galen P, Gan OI, Notta E, Brown AMK, et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* (2013) 339:543–8. doi:10.1126/science.1227670
85. Eisen M. *Mathematical Models in Cell Biology and Cancer Chemotherapy*. (Vol. 30). Berlin: Springer Science & Business Media (2013).
86. Phillips TM, McBride WH, Pajonk F. The response of CD24-/low/CD44+ breast cancer-initiating cells to radiation. *J Natl Cancer Inst* (2006) 98:1777–85. doi:10.1093/jnci/djj495
87. Kim JJ, Tannock IF. Repopulation of cancer cells during therapy: an important cause of treatment failure. *Nat Rev Cancer* (2005) 5:516–25. doi:10.1038/nrc1650
88. Lett JT. Damage to cellular DNA from particulate radiations, the efficacy of its processing and the radiosensitivity of mammalian cells. *Radiat Environ Biophys* (1992) 31:257–77. doi:10.1007/BF01210207
89. Coates J. Motivation for the inclusion of genetic risk factors of radiosensitivity alongside dosimetric and clinical parameters in predicting normal tissue effects. *Acta Oncol* (2015) 54:1230–1. doi:10.3109/0284186X.2014.999163
90. Jones B, Dale RG. *Radiobiology of High Dose Fractions*. *Stereotact. Body Radiother*. Berlin: Springer (2015). p. 67–86.
91. Guerrero M, Li XA. Extending the linear-quadratic model for large fraction doses pertinent to stereotactic radiotherapy. *Phys Med Biol* (2004) 49:4825–35. doi:10.1088/0031-9155/49/20/012
92. Takahashi W, Nakajima M, Yamamoto N, Tsuji H, Kamada T, Tsujii H. Carbon ion radiotherapy in a hypofractionation regimen for stage I non-small-cell lung cancer. *J Radiat Res* (2014) 55:i26–7. doi:10.1093/jrr/rrt216
93. Webb S, Nahum AE. A model for calculating tumour control probability in radiotherapy including the effects of inhomogeneous distributions of dose and clonogenic cell density. *Phys Med Biol* (1993) 38:653. doi:10.1088/0031-9155/38/6/001
94. Dale RG. Time-dependent tumour repopulation factors in linear-quadratic equations – implications for treatment strategies. *Radiother Oncol* (1989) 15:371–81. doi:10.1016/0167-8140(89)90084-4
95. Tilly N, Brahme A, Carlsson J, Glimelius B. Comparison of cell survival models for mixed LET radiation. *Int J Radiat Biol* (1999) 75:233–43. doi:10.1080/095530099140690
96. Ling CC, Chen CH, Fuks Z. An equation for the dose response of radiation-induced apoptosis: possible incorporation with the LQ model. *Radiother Oncol* (1994) 33:17–22. doi:10.1016/0167-8140(94)90081-7
97. Wang JZ, Huang Z, Lo SS, Yuh WTC, Mayr NA. A generalized linear-quadratic model for radiosurgery, stereotactic body radiation therapy, and high-dose rate brachytherapy. *Sci Transl Med* (2010) 2:39ra48. doi:10.1126/scitranslmed.3000864
98. Kellerer AM, Rossi HH. The theory of dual radiation action. *Curr Top Radiat Res* (1974) VIII:85–158.
99. Dale RG, Jones B. The assessment of RBE effects using the concept of biologically effective dose. *Int J Radiat Oncol Biol Phys* (1999) 43:639–45. doi:10.1016/S0360-3016(98)00364-2
100. Carabe-Fernandez A, Dale RG, Jones B. The incorporation of the concept of minimum RBE (RbEmin) into the linear-quadratic model and the potential for improved radiobiological analysis of high-LET treatments. *Int J Radiat Biol* (2007) 83:27–39. doi:10.1080/09553000601087176
101. Sachs RK, Hahnfeld P, Brenner DJ. Review the link between low-LET dose-response relations and the underlying kinetics of damage production/repair/misrepair. *Int J Radiat Biol* (1997) 72:351–74. doi:10.1080/095530097143149
102. Zaider M, Hanin L. Tumor control probability in radiation treatment. *Med Phys* (2011) 38:574–83. doi:10.1118/1.3521406
103. Gagliardi G, Bjöhle J, Lax I, Ottolenghi A, Eriksson F, Liedberg A, et al. Radiation pneumonitis after breast cancer irradiation: analysis of the complication probability using the relative seriality model. *Int J Radiat Oncol Biol Phys* (2000) 46:373–81. doi:10.1016/S0360-3016(99)00420-4
104. Boersma LJ, Damen EMF, De Boer RW, Muller SH, Olmos RAV, Van Zandwijk N, et al. Estimation of overall pulmonary function after irradiation using dose-effect relations for local functional injury. *Radiother Oncol* (1995) 36:15–23. doi:10.1016/0167-8140(95)01580-A
105. Rancati T, Fiorino C, Gagliardi G, Cattaneo GM, Sanguineti G, Borca VC, et al. Fitting late rectal bleeding data using different NTCP models: results from an Italian multi-centric study (AIROPROS0101). *Radiother Oncol* (2004) 73:21–32. doi:10.1016/j.radonc.2004.08.013
106. Kutcher GJ, Burman C. Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method. *Int J Radiat Oncol Biol Phys* (1989) 16:1623–30. doi:10.1016/0360-3016(89)90972-3
107. Stavrev P, Stavreva N, Niemierko A, Goitein M. Generalization of a model of tissue response to radiation based on the idea of functional subunits and binomial statistics. *Phys Med Biol* (2001) 46:1501–18. doi:10.1088/0031-9155/46/5/312
108. Niemierko A, Goitein M. Modeling of normal tissue response to radiation: the critical volume model. *Int J Radiat Oncol Biol Phys* (1993) 25:135–45. doi:10.1016/0360-3016(93)90156-P
109. Stavrev P, Stavreva N, Sharplin J, Fallone BG, Franko A. Critical volume model analysis of lung complication data from different strains of mice. *Int J Radiat Biol* (2005) 81:77–88. doi:10.1080/09553000400027910
110. Yaes RJ, Kalend A. Local stem cell depletion model for radiation myelitis. *Int J Radiat Oncol Biol Phys* (1988) 14:1247–59. doi:10.1016/0895-7177(88)90651-6
111. Jackson A, Ten Haken RK, Robertson JM, Kessler ML, Kutcher GJ, Lawrence TS. Analysis of clinical complication data for radiation hepatitis using a parallel architecture model. *Int J Radiat Oncol Biol Phys* (1995) 31:883–91. doi:10.1016/0360-3016(94)00471-4
112. Hartford AC, Niemierko A, Adams JA, Urie MM, Shipley WU. Conformal irradiation of the prostate: estimating long-term rectal bleeding risk using dose-volume histograms. *Int J Radiat Oncol Biol Phys* (1996) 36:721–30. doi:10.1016/S0360-3016(96)00366-5
113. Lebesque JV, Bruce AM, Kroes APG, Touw A, Shouman T, Van Herk M. Variation in volumes, dose-volume histograms, and estimated normal tissue complication probabilities of rectum and bladder during conformal radiotherapy of T3 prostate cancer. *Int J Radiat Oncol Biol Phys* (1995) 33:1109–19. doi:10.1016/0360-3016(95)00253-7
114. Adamus-Górka M, Mavroidis P, Lind BK, Brahme A. Comparison of dose response models for predicting normal tissue complications from cancer radiotherapy: application in rat spinal cord. *Cancers (Basel)* (2011) 3:2421–43. doi:10.3390/cancers3022421
115. Tomczak J. Prediction of breast cancer recurrence using classification restricted boltzmann machine with dropping. *arXiv* (2013) 13086324:1–9.
116. Lee S, Ybarra N, Jeyaseelan K, Seuntjens J, El Naqa I, Faria S, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. *Med Phys* (2015) 42:2421–30. doi:10.1118/1.4915284
117. van der Schaaf A, Langendijk JA, Fiorino C, Rancati T. Embracing phenomenological approaches to normal tissue complication probability modeling: a question of method. *Int J Radiat Oncol Biol Phys* (2015) 91:468–71. doi:10.1016/j.ijrobp.2014.10.017
118. van der Schaaf A, Xu C-J, van Luijk P, van't Veld AA, Langendijk JA, Schilstra C. Multivariate modeling of complications with data driven variable selection:

- guarding against overfitting and effects of data set size. *Radiother Oncol* (2012) 105:115–21. doi:10.1016/j.radonc.2011.12.006
119. Anagnostou T, Remzi M, Lykourinas M, Djavan B. Artificial neural networks for decision-making in urologic oncology. *Eur Urol* (2003) 43:596–603. doi:10.1016/S0302-2838(03)00133-7
  120. Djavan B, Remzi M, Zlotta A, Seitz C, Snow P, Marberger M. Novel artificial neural network for early detection of prostate cancer. *J Clin Oncol* (2002) 20:921–9. doi:10.1200/JCO.20.4.921
  121. Jerez-Aragónés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artif Intell Med* (2003) 27:45–63. doi:10.1016/S0933-3657(02)00086-6
  122. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* (2000) 19:541–61. doi:10.1002/(SICI)1097-0258(20000229)19:4<541:AID-SIM355>3.0.CO;2-V
  123. Tomatis S, Rancati T, Fiorino C, Vavassori V, Fellin G, Cagna E, et al. Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model. *Phys Med Biol* (2012) 57:1399–412. doi:10.1088/0031-9155/57/5/1399
  124. Whitley D, Starkweather T, Bogart C. Genetic algorithms and neural networks: optimizing connections and connectivity. *Parallel Comput* (1990) 14:347–61. doi:10.1016/0167-8191(90)90086-O
  125. Belew RK, McInerney J, Schraudolph NN. Evolving networks: using the genetic algorithm with connectionist learning. *Artif Life II* (1992) 10:511–47.
  126. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. *Med Phys* (2005) 32:318–25. doi:10.1118/1.1835611
  127. Chen S, Zhou S, Zhang J, Yin F-F, Marks LB, Das SK. A neural network model to predict lung radiation-induced pneumonitis. *Med Phys* (2007) 34:3420–7. doi:10.1118/1.2759601
  128. Specht DF. A general regression neural network. *IEEE Trans Neural Networks* (1991) 2:568–76. doi:10.1109/72.97934
  129. Blanco AI, Chao KSC, El Naqa I, Franklin GE, Zakarian K, Vivic M, et al. Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. *Int J Radiat Oncol Biol Phys* (2005) 62:1055–69. doi:10.1016/j.ijrobp.2004.12.076
  130. Hoffmann H. Kernel PCA for novelty detection. *Pattern Recognit* (2007) 40:863–74. doi:10.1016/j.patcog.2006.07.009
  131. Mika S, Schölkopf B, Smola A, Müller K, Scholz M, Rätsch G. Kernel PCA and de-noising in feature spaces. *Analysis* (1999) 11:536–42.
  132. Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud J-D, editors. *Artif. Neural Networks – ICANN'97 SE – 93*. (Vol. 1327), Berlin: Springer (1997). p. 583–8.
  133. El Naqa I. Outcomes modeling (Chapter 24). In: Starkschall G, Siochi C, editors. *Informatics in Radiation Oncology*. Boca Raton, FL (2013). p. 257–76.
  134. Kitano H. Systems biology: a brief overview. *Science* (2002) 295:1662–4. doi:10.1126/science.1069492
  135. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine introduction: paradigm changes in health care. *J Proteome Res* (2004) 3(2):179–96. doi:10.1021/pr0499693
  136. Laubenbacher R, Hower V, Jarrah A, Torti SV, Shulaev V, Mendes P, et al. A systems biology view of cancer. *Biochim Biophys Acta Rev Cancer* (2009) 1796:129–39. doi:10.1016/j.bbcan.2009.06.001
  137. Feinendegen L, Hahnfeldt P, Schadt EE, Stumpf M, Voit EO. Systems biology and its potential role in radiobiology. *Radiat Environ Biophys* (2007) 47:5–23. doi:10.1007/s00411-007-0146-8
  138. Unger K. Integrative radiation systems biology. *Radiat Oncol* (2014) 9:21. doi:10.1186/1748-717X-9-21
  139. Smith WP, Doctor J, Meyer J, Kalet IJ, Phillips MH. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model. *Artif Intell Med* (2009) 46:119–30. doi:10.1016/j.artmed.2008.12.002
  140. Oh JH, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol* (2011) 56:1635–51. doi:10.1088/0031-9155/56/6/008
  141. Messer JA, Mohamed ASR, Hutcheson KA, Ding Y, Lewin JS, Wang J, et al. Magnetic resonance imaging of swallowing-related structures in nasopharyngeal carcinoma patients receiving IMRT: longitudinal dose-response characterization of quantitative signal kinetics. *Radiother Oncol* (2016) 118:315–22. doi:10.1016/j.radonc.2016.01.011
  142. El Naqa I, Suneja G, Lindsay PE, Hope AJ, Alaly JR, Vivic M, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys Med Biol* (2006) 51:5719–35. doi:10.1088/0031-9155/51/22/001
  143. Aho K, Derryberry D, Peterson T. Model selection for ecologists: the world-views of AIC and BIC. *Ecology* (2014) 95:631–6. doi:10.1890/13-1452.1
  144. Akaike H. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado (1973). p. 267–81.
  145. Schwarz G. Estimating the dimension of a model. *Ann Stat* (1978) 6:461–4. doi:10.1214/aos/1176344136
  146. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* (1995) 90(430):773–95.
  147. Coates J, Jeyaseelan AK, Ybarra N, Tao J, David M, Faria S, et al. Evaluation and visualization of radiogenomic modeling frameworks for the prediction of normal tissue toxicities. In: Jaffray DA, editor. *World Congr. Med. Phys. Biomed. Eng. June 7-12, 2015, Toronto, Canada*. Berlin: Springer International Publishing (2015). p. 517–20.
  148. Deasy JO, Bentzen SM, Jackson A, Ten Haken RK, Yorke ED, Constine LS, et al. Improving normal tissue complication probability models: the need to adopt a “data-pooling” culture. *Int J Radiat Oncol Biol Phys* (2010) 76:S151–4. doi:10.1016/j.ijrobp.2009.06.094
  149. Li XA, Alber M, Deasy JO, Jackson A, Jee K-WK, Marks LB, et al. The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM. *Med Phys* (2012) 39:1386–409. doi:10.1118/1.3685447
  150. Sanchez-Nieto B, Nahum AE. BIOPLAN: software for the biological evaluation of Radiotherapy treatment plans. *Med Dosim* (2000) 25:71–6. doi:10.1016/S0958-3947(00)00031-5
  151. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. *Med Phys* (2003) 30:979–85. doi:10.1118/1.1568978
  152. Zietman A. Proton beam and prostate cancer: an evolving debate. *Reports Pract Oncol Radiother* (2013) 18:338–42. doi:10.1016/j.rpor.2013.06.001
  153. Paganetti H, Niemierko A, Ancukiewicz M, Gerweck LE, Goitein M, Loeffler JS, et al. Relative biological effectiveness (RBE) values for proton beam therapy. *Int J Radiat Oncol Biol Phys* (2002) 53:407–21. doi:10.1016/S0360-3016(02)02754-2
  154. Zietman AL, Bae K, Slater JD, Shipley WU, Efstathiou JA, Coen JJ, et al. Randomized trial comparing conventional-dose with high-dose conformal radiation therapy in early-stage adenocarcinoma of the prostate: long-term results from proton radiation oncology group/American College of Radiology 95-09. *J Clin Oncol* (2010) 28:1106–11. doi:10.1200/JCO.2009.25.8475
  155. Ishikawa H, Tsuji H, Kamada T, Yanagi T, Mizoe J-E, Kanai T, et al. Carbon ion radiation therapy for prostate cancer: results of a prospective phase II study. *Radiother Oncol* (2006) 81:57–64. doi:10.1016/j.radonc.2006.08.015
  156. Tsuji H, Yanagi T, Ishikawa H, Kamada T, Mizoe J, Kanai T, et al. Hypofractionated radiotherapy with carbon ion beams for prostate cancer. *Int J Radiat Oncol Biol Phys* (2005) 63:1153–60. doi:10.1016/j.ijrobp.2005.04.022
  157. Slater JD, Rossi CJ, Yonemoto LT, Bush DA, Jabola BR, Levy RP, et al. Proton therapy for prostate cancer: the initial Loma Linda University experience. *Int J Radiat Oncol Biol Phys* (2004) 59:348–52. doi:10.1016/j.ijrobp.2003.10.011
  158. Talcott JA, Rossi C, Shipley WU, Clark JA, Slater JD, Niemierko A, et al. Patient-reported long-term outcomes after conventional and high-dose combined proton and photon radiation for early prostate cancer. *J Urol* (2010) 184:1993. doi:10.1016/j.juro.2010.07.024
  159. Fontana AO, Augsburg MA, Grosse N, Guckenberger M, Lomax AJ, Sartori AA, et al. Differential DNA repair pathway choice in cancer cells after proton- and photon-irradiation. *Radiother Oncol* (2015) 116(3):374–80. doi:10.1016/j.radonc.2015.08.014
  160. Gerelchuluun A, Manabe E, Ishikawa T, Sun L, Itoh K, Sakae T, et al. The major DNA repair pathway after both proton and carbon-ion radiation is NHEJ, but the HR pathway is more relevant in carbon ions. *Radiat Res* (2015) 183:345–56. doi:10.1667/RR13904.1



161. Tomczak JM. Application of classification restricted Boltzmann machine to medical domains. *International Conference on Innovative Trends in Science, Engineering and Management (ICITSEM 2014)*. Dubai, UAE: (2014).
162. Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, et al. *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014, Held in Conjunction with MICCAI 2014*. Cambridge, MA: Springer (2014). [Revised Selected Papers. vol. 8848].
163. Kozioł JA, Tan EM, Dai L, Ren P, Zhang J-Y. Restricted Boltzmann machines for classification of hepatocellular carcinoma. *Comput Biol J* (2014) 2014:5. doi:10.1155/2014/418069
164. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* (2006) 313:504–7. doi:10.1126/science.1127647
165. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. *Comput. Robot Vis. (CRV), 2015 12th Conf., IEEE*. Halifax, NS (2015). p. 133–8.
166. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans Comput Biol Bioinform* (2015) 12:928–37. doi:10.1109/TCBB.2014.2377729

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Coates, Souhami and El Naqa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.