IISE Transactions

Design & Manufacturing

Quality & Reliability Engineering

IISE Transactions

ISSN: 2472-5854 (Print) 2472-5862 (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Integration of biological and statistical models toward personalized radiation therapy of cancer

Xiaonan Liu, Mirek Fatyga, Teresa Wu & Jing Li

To cite this article: Xiaonan Liu, Mirek Fatyga, Teresa Wu & Jing Li (2019) Integration of biological and statistical models toward personalized radiation therapy of cancer, IISE Transactions, 51:3, 311-321, DOI: 10.1080/24725854.2018.1486054

To link to this article: https://doi.org/10.1080/24725854.2018.1486054

IISE TRANSACTIONS
2019, VOL. 51, NO. 3, 311–321
https://doi.org/10.1080/24725854.2018.1486054

Integration of biological and statistical models toward personalized radiation therapy of cancer

Xiaonan Liu[a], Mirek Fatyga[b], Teresa Wu[a], and Jing Li[a]

[a]School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA; [b]Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix, AZ, USA

ABSTRACT

Radiation Therapy (RT) is one of the most common treatments for cancer. To understand the impact of radiation toxicity on normal tissue, a Normal Tissue Complication Probability (NTCP) model is needed to link RT dose with radiation-induced complications. There are two types of NTCP models: biological and statistical models. Biological models have good generalizability but low accuracy, as they cannot factor in patient-specific information. Statistical models can incorporate patient-specific variables, but may not generalize well across different studies. We propose an integrated model that borrows strength from both biological and statistical models. Specifically, we propose a novel model formulation followed by an efficient parameter estimation algorithm, and investigate statistical properties of the estimator. We apply the integrated model to a real dataset of prostate cancer patients treated with Intensity Modulated RT at the Mayo Clinic Arizona, who are at risk of developing the grade 2+ acute rectal complication. The integrated model achieves an Area Under the Curve (AUC) level of 0.82 in prediction, whereas the AUCs for the biological and statistical models are only 0.66 and 0.76, respectively. The superior performance of the integrated model is also consistently observed over different simulation experiments.

ARTICLE HISTORY
Received 23 December 2016
Accepted 14 May 2018

KEYWORDS
Model integration; classification; radiation toxicity

1. Introduction

Radiation Therapy (RT) is one of the most common treatments for cancer, either by itself or combined with other forms of treatments. Although the goal of RT is tumor control, it is also important in RT planning to spare the normal (surrounding) tissue from radiation toxicity that could greatly affect a patient's quality of life. To understand the impact of radiation toxicity on normal tissue, a model is needed to link radiation dose of the RT with radiation-induced complications. This is known as the Normal Tissue Complication Probability (NTCP) model. There are two types of NTCP models: biological models and statistical models. Biological models are built on the understanding of normal tissue cells' response to injury by ionizing radiation. Typical works include the Lyman model (Lyman, 1985), the Lyman-Kutcher-Burman (LKB) model (Deasy, 2000), the generalized Lyman model (Tucker *et al.*, 2008), and the relative seriality model (Källman *et al.*, 1992). Statistical models aim to associate features of radiation dose distribution among the receiving tissue with the risk of developing certain complications. There are many statistical models to choose from, since this is a typical classification/prediction problem in statistics, so that theoretically speaking, any classification/prediction algorithm can be a potential candidate. Statistical models have been popularly used in recent years for NTCP modeling (Liu *et al.*, 2010; Michalski *et al.*, 2010; Boomsma *et al.*, 2012; Gulliford *et al.*, 2012; Cella *et al.*, 2013).

Next, we will discuss the advantages and disadvantages of biological and statistical models. One of the most important advantages for biological models is that the results can be generalized beyond a particular study; this is due to biological models being built on radiobiological principles. This provides convenience for clinical utilization, as clinicians may refer to published results when they do not have the data to build a biological model for their specific practice. However, current knowledge about the biological reaction to radiation is still limited. As a result, biological models usually have a low accuracy in predicting the risk of radiation-induced complications. Another major contributor to the low prediction accuracy is the lack of "personalization," i.e., biological models do not factor in patient-specific information when linking radiation dose with complications. However, mounting evidence has shown that even with the same radiation dose distribution, different patients have different susceptibility to radiation toxicity, and thereby different risks of developing complications (Boomsma *et al.*, 2012; Cella *et al.*, 2013).

In contrast, statistical models have the flexibility of incorporating patient-specific variables in addition to radiation dose-related variables as predictors. In fact, modern developments in statistics, especially machine learning, makes it possible to include a large collection of patient-specific variables, such as demographics, health conditions, and even genetic and epigenetic markers. This enables a truly

CONTACT Jing Li jinglz@asu.edu
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uiie.

personalized approach in NTCP modeling. In addition to personalization, statistical models can better account for the special RT data characteristics. One distinct characteristic of the RT data is that the sample fraction of a complication (i.e., the fraction of patients with the complication in the sample dataset used for statistical modeling) is typically much larger than the population fraction of the complication. This is because the latter fraction is usually a very small number (otherwise, the RT would not be permitted for practice). If the same fraction were used in the sample dataset, there would be too few patients with the complication, making the dataset uninformative. Increasing the sample fraction of the complication creates richer information content in the dataset, but the resulting statistical model may be biased and inconsistent. Fortunately, theoretical analysis is possible to quantify the level of inconsistency and bias for most statistical models. The result can further guide the development of effective consistency and bias correction strategies. Due to the aforementioned advantages, statistical models usually have better prediction accuracies for complications than biological models. However, a major drawback of statistical models is that the results heavily depend on the particular dataset used in each study, so they may not generalize well.

In this article, we propose a general framework with specific methods for biological and statistical model integration in NTCP modeling. Our goal is to preserve the biological knowledge in NTCP modeling, and meanwhile use statistical modeling strategies to allow for inclusion of patient-specific variables and to better account for the special RT data characteristic. The integrated approach is expected to have better generalizability and predictive power than using biological and statistical models in isolation.

The contributions of this paper are two-fold:

1. Novel model development: We propose the first-of-its-kind framework for biological and statistical model integration. Under the framework, we propose the details for developing the integrated model, including a novel model formulation, an efficient algorithm for parameter estimation, and theoretical analysis guided consistency and bias corrections to guarantee that the model has good statistical properties.
2. Real-data application: We apply the integrated model to a dataset of prostate cancer patients treated with Intensity-Modulated RT (IMRT), an advanced type of RT, at the Mayo Clinic Arizona. These patients are at risk of developing a serious complication called the grade 2+ acute rectal complication with symptoms including anal pain, diarrhea, and rectal obstruction. The integrated model achieves higher accuracy in predicting the complication compared with the statistical and biological models used separately. Also, we perform extensive simulation studies on virtual patients whose data are sampled from the distribution of the real patients. Under various simulation settings, the integrated model outperforms the statistical and biological models in prediction accuracy, due to the inclusion of patient-specific variables and in generalizability across different datasets due to the consideration of radiobiological principles.

The rest of this article is organized as follows: Section 2 presents the development of the integrated model; Section 3 presents the application and simulation studies. Section 4 concludes the paper.

2. Integration of biological and statistical models in NTCP modeling

We propose a model integration framework that includes three major steps:

1. One should start with an in-depth understanding of the biological model, especially the biological meanings of the model parameters. This guides the decision on which parameter(s) are appropriate to be personalized.
2. Then, the selected biological model parameter(s) is (are) linked with patient-specific variables in an appropriate way, producing an integrated model formulation. This formulation needs to be properly designed to account for the potential high-dimensionality of patient-specific variables and biological constraints on the model parameters. Under the formulation, an optimization algorithm is further developed to estimate the parameters of the integrated model, with considerations on optimality and efficiency.
3. Finally, statistical properties of the integrated model, such as consistency and bias, are investigated, and corrections are made in order to produce consistent and unbiased estimators for the model parameters.

In what follows, we present the details of the proposed approaches for accomplishing steps 1 to 3 in sub-sections 2.1 to 2.3, respectively.

2.1. Understanding the biological model in NTCP modeling

The goal of NTCP modeling is to link the radiation dose delivered to a normal tissue/organ (not the organ with cancer) with the risk/probability that the normal organ will develop a complication. In this article, we focus on the rectum, which is a normal organ that could develop a complication from radiation toxicity, for patients with prostate cancer. For each patient, treatment planning software generates a three-dimensional (3-D) dose map, which contains the radiation dose value on each voxel of a 3-D image (CT or MRI) of the rectum. In biological NTCP models, the 3-D dose map is first converted to a Dose-Volume-Histogram (DVH) by binning the voxel-wise dose values. Figure 1 shows an example of a DVH, in which the horizontal axis corresponds to dose values from low to high; the vertical axis is the fraction of voxels (i.e., volume) of the rectum that receive a certain dose.

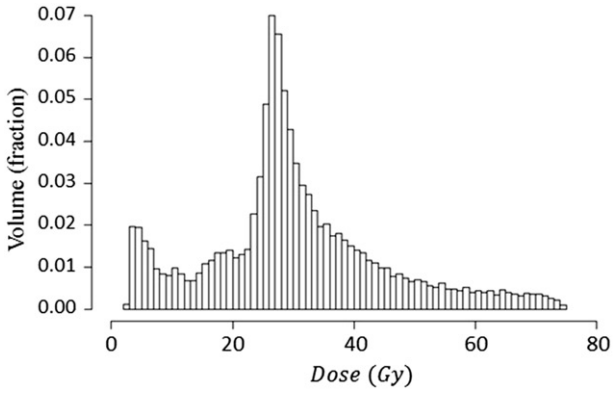

Figure 1. DVH of the rectum of a prostate cancer patient receiving IMRT.

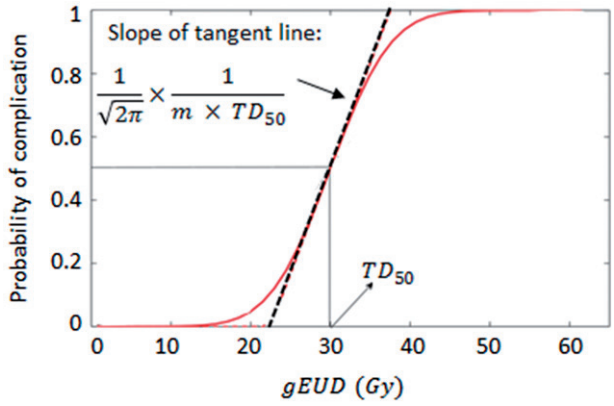

Figure 2. Probability of complication, $P(Y = 1)$, with respect to $gEUD$ in the LKB model.

In biological NTCP models, radiobiological principles are typically used to guide: (i) the selection or development of a metric from DVH, which captures the essence of normal tissue cells' biological response to injury by ionizing radiation; and (ii) the determination of the functional form of the relationship between the metric and the probability of developing the complication. Next, we will present the details of a well-known biological NTCP model, called the LKB model (Gulliford *et al.*, 2012). In the LKB model, a specific metric of DVH is used, called the generalized Equivalent Uniform Dose (gEUD), which takes the form of Equation (1):

$$gEUD = \left(\sum\nolimits_i v_i D_i^{1/n}\right)^n. \tag{1}$$

Here v_i is the fractional tissue volume receiving a dose D_i at the ith DVH bin. v_i and D_i can be readily obtained from a given DVH. n is a parameter of the LKB model, the meaning of which will be discussed later. It can be seen from Equation (1) that $gEUD$ collapses the complex dose distribution represented by a DVH into a single metric. The biological rationale behind this specific form of collapsing is that $gEUD$ represents the uniform dose that, if delivered over the same number of fractions as the real but non-uniform dose distribution, yields the same radiobiological effect (Li *et al.*, 2012). Furthermore, the LKB model links $gEUD$ with the probability for a patient to develop the complication by a sigmoid-shape function, that is

$$P(Y = 1) = \phi\left(\frac{gEUD - TD_{50}}{m \times TD_{50}}\right). \tag{2}$$

Here, Y is an indicator variable; $Y = 1$ represents that the patient has the complication and $Y = 0$ otherwise. $\phi(\cdot)$ is the cumulative probability function for the standard normal distribution. TD_{50} and m are two other LKB model parameters. Figure 2 shows the function of $P(Y = 1)$ with respect to $gEUD$.

Next, we will discuss the meanings of the three LKB model parameters, TD_{50}, m, and n. According to Equation (2), TD_{50} is the $gEUD$ given to the normal tissue that results in 50% probability of the complication (Figure 2). Intuitively, TD_{50} can be considered to be the tolerance dose for developing the complication. That is, a slightly higher $gEUD$ than TD_{50} will make the patient have a risk of developing the complication that is greater than a random guess. Furthermore, to understand parameter m, we take the partial derivative of $P(Y = 1)$ with respect to $gEUD$ and evaluate this partial derivative at $gEUD = TD_{50}$, which gives:

$$\left|\frac{\partial\ P(Y=1)}{\partial\ gEUD}\right|_{gEUD=TD_{50}} = \frac{1}{\sqrt{2\pi}} \times \frac{1}{m \times TD_{50}}. \tag{3}$$

Equation (3) means that:

$$\frac{1}{\sqrt{2\pi}} \times \frac{1}{m \times\ TD_{50}}$$

is the slope of the "$P(Y = 1)$ vs $gEUD$" curve at $gEUD = TD_{50}$ (Figure 2). The slope of this curve reflects the sensitivity of the complication probability with respect to a change in $gEUD$. The bigger the slope, i.e., the smaller the $m\ * TD_{50}$, the higher the sensitivity. Therefore, the meaning of m is that it is inversely related to the sensitivity of the complication probability with a fixed TD_{50}. Finally, we discuss parameter n. In the LKB model, n is constrained to be between zero and one. When $n = 1$, it is obvious from Equation (1) that $gEUD$ becomes the average dose received by the normal tissue. When $n = 0$, Proposition 1 shows that $gEUD$ becomes the maximum dose received by the tissue (please see the proof in Appendix A).

Proposition 1.

$$\lim_{n\to 0}\ gEUD = \lim_{n\to 0}\left(\sum\nolimits_i v_i D_i^{1/n}\right)^n = \max_i D_i.$$

To obtain the parameters of the LKB model in a particular clinical study, we need a collection of patient data, based on which the parameters can be estimated by a Maximum Likelihood Estimation (MLE) approach. The LKB model has been extensively used to model various types of complications on many normal tissues/organs for major modern RT techniques. To consolidate the results from various studies, the American Association of Physicists in Medicine and the American Society for Radiation Oncology jointly funded a multidisciplinary study, called Quantitative Analysis of Normal Tissue Effects in the Clinic, in order to summarize the existing findings and develop clinical guidance. Reference values for the LKB model parameters were provided for clinically significant complications of 16 tissues/organs (Bentzen *et al.*, 2010).

2.2. Integration of patient-specific information into the biological model

2.2.1. Formulation

Based on the understanding of the LKB model parameters in the previous section, we now discuss which parameter(s) are more appropriate to be personalized. According to the definition, TD_{50} reflects the radiation dose a patient can tolerate before developing a complication with a chance higher than a random guess. Previous studies have shown that different patients have different tolerances, depending on their age, health conditions, and even genetics (Boomsma *et al.*, 2012; Cella *et al.*, 2013). Therefore, it is obvious that TD_{50} should be patient-specific. Furthermore, $m * TD_{50}$ reflects the sensitivity of complication probability with respect to a change in radiation dose. There is no clear medical evidence so far to support whether or not this sensitivity should be patient-specific. Therefore, we choose not to personalize $m * TD_{50}$. Finally, regarding parameter n, there is solid evidence in the literature to suggest that it is more organ-specific than patient-specific (Gulliford *et al.*, 2012; Li *et al.*, 2012). Specifically, in the RT literature, organs are classified into parallel organs (e.g., lung, kidney, and liver) and serial organs (e.g., spinal cord, intestines, and optic nerves). Sub-units of a parallel organ function relatively independently, so radiation damage to a small region does not make the whole organ dysfunctional. Therefore, the probability of developing a complication for a parallel organ should be more closely related to the average dose it receives, i.e., with $n \to 1$. On the other hand, a serial organ tends to exhibit the complication if one sub-unit is incapacitated, so that the probability of developing a complication for a serial organ is more related to the maximum dose, i.e., with $n \to 0$.

Due to these considerations, we propose to incorporate patient-specific variables into TD_{50}. Let $x_{1j}, \ldots, x_{pj}$ be p patient-specific variables for patient j. Then,

$$TD_{50j} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{kj}, \tag{4}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are parameters to be estimated. Also, because we have decided not to personalize the sensitivity, we can replace the $m * TD_{50}$ in Equation (2) by a new parameter σ. Therefore, Equation (2) becomes:

$$P(y_j = 1) = \phi\left(\frac{gEUD_j(n) - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{kj}\right)}{\sigma}\right). \tag{5}$$

Let $\alpha_{p+1} \triangleq \sigma^{-1}$, $\alpha_0 \triangleq -\beta_0\sigma^{-1}$, and $\alpha_k \triangleq -\beta_k\sigma^{-1}$. Then, Equation (5) becomes:

$$P(y_j = 1) = \phi\left(\alpha_0 + \sum_{k=1}^{p} \alpha_k x_{kj} + \alpha_{p+1} gEUD_j(n)\right). \tag{6}$$

Furthermore, using Equation (6), the log-likelihood function can be written as

$$l(\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1}, n) = \sum_j \{y_j \log P(y_j = 1) + (1 - y_j)\log P(y_j = 0)\}. \tag{7}$$

Note that there can be a relatively large number of patient-specific variables to be included in the model, while the sample size in this type of studies is usually limited. This makes the estimation of model parameters unstable. To address this challenge, we follow a similar idea to the LASSO model (Tibshirani, 1996) and add an l_1-penalty to the model parameters, which results in an optimization problem as follows:

$$\min_{\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1},\ n} \left\{-l(\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1}, n) + \lambda \sum_{k=1}^{p} \alpha_k, \right.$$
$$s.t.\ \alpha_{p+1} \geq 0. \tag{8}$$

In Equation (8), λ is a tuning parameter. The constraint of $\alpha_{p+1} \geq 0$ is to satisfy the biological validity that a radiation dose always has a non-negative effect on the risk of developing a complication.

2.2.2. Estimation

The optimization problem in Equation (8) is difficult to solve, due to the complicated relationship between the objective function and the parameter n. On the other hand, with a fixed n, $gEUD_j(n)$ can be computed from the DVH of each patient, and consequently Equation (8) becomes a convex optimization problem with a linear constraint that is much easier to solve. This motivates us to treat n as a tuning parameter rather than a parameter to be directly optimized. As a result, Equation (8) can be written as Equation (9):

$$\min_{\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1}} \left\{-l_n(\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1}) + \lambda \sum_{k=1}^{p} \alpha_k, \right.$$
$$s.t.\ \alpha_{p+1} \geq 0. \tag{9}$$

To solve Equation (9), we propose an efficient algorithm based on the result of Proposition 2. The algorithm works by first solving the unconstrained optimization, which can be done by an efficient convex solver, and then using a simple fix to obtain the solution to the constrained optimization. Please see the proof of Proposition 2 in Appendix B. The optimal tuning parameters, λ^* and n^* can be found by a grid search based on a model selection criterion such as the AIC or BIC. The range of λ, $[\lambda_{min}, \lambda_{max}]$, is chosen such that λ_{max} results in no patient-specific variables being selected (i.e., the sparsest model), and λ_{min} results in all patient-specific variables or the number of patient-specific variables equal to the sample size being selected, whichever is smaller (i.e., the statistically plausible densest model). To set the range for n, we can run the original LKB model and obtain the confidence interval for n. This confidence internal can be used as $[n_{min}, n_{max}]$.

Proposition 2. *Let $\hat{\alpha}_0, \ldots, \hat{\alpha}_{p+1}$ be the solution to the optimization in Equation (9). Let $\alpha_0^*, \ldots, \alpha_{p+1}^*$ be the solution to unconstrained problem. If $\alpha_{p+1}^* \geq 0$, then $(\hat{\alpha}_0, \ldots, \hat{\alpha}_{p+1}) = (\alpha_0^*, \ldots, \alpha_{p+1}^*)$. If $\alpha_{p+1}^* < 0$, then $(\hat{\alpha}_0, \ldots, \hat{\alpha}_p) = (\alpha_0^*, \ldots, \alpha_p^*)$ and $\hat{\alpha}_{p+1} = 0$.*

2.3. Consistency and bias correction for the integrated model

After the solution to Equation (9), i.e., $\hat{\alpha}_0, \ldots, \hat{\alpha}_{p+1}$, is obtained, we will use the zero and non-zero patterns in

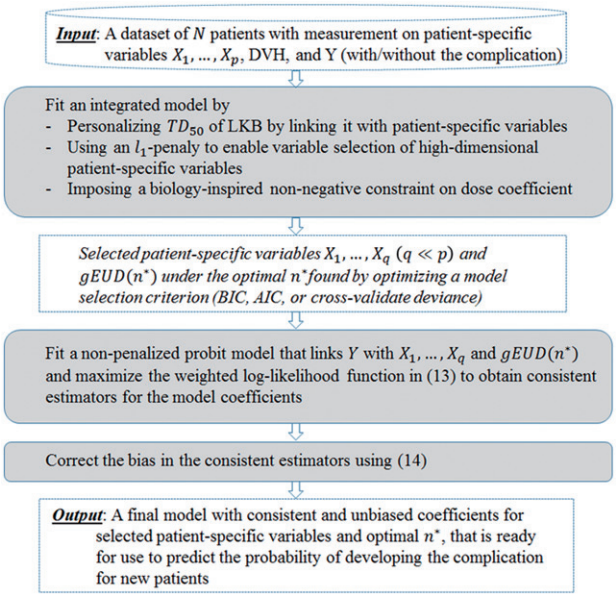

Figure 3. A flow chart of the steps for the proposed integrated model.

$\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ to select patient-specific variables. The selected variables will be used together with $gEUD_j(n^*)$ to fit a non-penalized model. This model can be used to predict the probability of developing the complication for new patients. Let $\widetilde{\alpha}_0, \ldots, \widetilde{\alpha}_q$ denote the estimated coefficients from this non-penalized model by MLE, that is

$$\begin{aligned}(\widetilde{\alpha}_0, \ldots, \widetilde{\alpha}_q) &= \max_{\alpha_0, \alpha_1, \ldots, \alpha_q} l_{n^*}(\alpha_0, \alpha_1, \ldots, \alpha_q) \\ &= \max_{\alpha_0, \alpha_1, \ldots, \alpha_q} \sum_j \{y_j \log P_{n^*}(y_j = 1) + (1 - y_j) \log P_{n^*}(y_j = 0)\},\end{aligned} \tag{10}$$

where

$$P_{n^*}(y_j = 1) = \phi\left(\alpha_0 + \sum_{k=1}^{q-1} \alpha_k x_{kj} + \alpha_q gEUD_j(n^*)\right). \tag{11}$$

The form of Equation (11) is known as the probit model. The rationale for re-fitting a non-penalized model is that the l_1-penalty is known to have a shrinking effect, which makes an l_1-penalized model a good variable selection model, but not necessarily a good predictive model (Hastie *et al.*, 2015). Next, we discuss two important statistical properties of the estimators $\widetilde{\alpha}_0, \ldots, \widetilde{\alpha}_q$, i.e., consistency and bias. In statistics, a consistent estimator is one that converges in probability to the true value of the parameter being estimated as the sample size goes to infinity. The bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated. If the bias is zero, the corresponding estimator is called an unbiased estimator. A good estimator should be consistent and unbiased.

2.3.1. Consistency

The population of patients going through a RT is typically heavy imbalanced, i.e., a very small fraction of the population will develop the complication of interest whereas the large majority will not. Otherwise, the RT would not have been permitted for practice. When a dataset is sampled from the population for NTCP modeling, a common strategy is to include a larger fraction of patients with the complication in the dataset than what this faction truly is in the population. This is to make sure that the dataset has enough samples with $Y = 1$ (i.e., with the complication) and thus rendering a meaningful statistical analysis (King and Zeng, 2001). However, when this sampling strategy is used, $\widetilde{\alpha}_0$ is not a statistically consistent estimator for α_0, although $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_q$ still are. This finding is summarized in Proposition 3.

Proposition 3. *Let τ and $\bar{y}$ be the fractions of patients with $Y = 1$ in the population and in the sample dataset, respectively. If $\bar{y} \neq \tau$, then $\widetilde{\alpha}_0$ is not a consistent estimator for α_0 while $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_q$ are consistent estimators for $\alpha_1, \ldots, \alpha_q$.*

To produce a consistent estimator, we propose a Maximum Weighted Likelihood Estimation (MWLE) to estimate the non-penalized model in Equation (11). In the MWLE, the log-likelihood function takes a weighted form, that is

$$\begin{aligned}l^w_{n^*}(\alpha_0, \alpha_1, \ldots, \alpha_q) = w_1 &\sum_{\{y_j=1\}} \log P_{n^*}(y_j = 1) \\ &+ w_0 \sum_{\{y_j=0\}} \log P_{n^*}(y_j = 0),\end{aligned} \tag{12}$$

where $w_1 = \tau/\bar{y}$, $w_0 = (1-\tau)/(1-\bar{y})$. By maximizing Equation (12), we can obtain estimates for $\alpha_0, \alpha_1, \ldots, \alpha_q$, denoted by $\breve{\alpha}_0, \ldots, \breve{\alpha}_q$. Proposition 4 shows that $\breve{\alpha}_0, \ldots, \breve{\alpha}_q$ are consistent estimators. Proofs of Propositions 3 and 4 share a similar idea to the proof of consistency for the weighted exogenous sampling maximum likelihood estimator in Manski and Lerman, (1977), and thus are skipped here due to space limits.

Proposition 4. *$\breve{\alpha}_0, \ldots, \breve{\alpha}_q$ are consistent estimators for $\alpha_0, \alpha_1, \ldots, \alpha_q$.*

Note that the weighted log-likelihood function proposed in Equation (12) assumes a known τ. In fact, τ is straightforward to obtain. For example, it can be estimated from a data source that only records patients' complications, such as the Electronic Health Records. As such a data source does not include the RT dose, it can be easily created and thus including a large patient population to grant an accurate estimate for τ. τ may also be obtained from published epidemiologic studies on the RT, which usually report the fraction of people developing the complication in a large population.

2.3.2. Bias

Although $\breve{\alpha}_0, \ldots, \breve{\alpha}_q$ are consistent estimators, they are still biased. Proposition 5 derives the bias of these estimators. Please see the proof in Appendix C.

Proposition 5. *Let $\breve{\boldsymbol{\alpha}} = (\breve{\alpha}_0, \ldots, \breve{\alpha}_q)$. Denote the true values of the parameters being estimated by $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_q)$. The bias of $\breve{\boldsymbol{\alpha}}$ is*

$$bias(\breve{\boldsymbol{\alpha}}) = E(\breve{\boldsymbol{\alpha}}) - \boldsymbol{\alpha} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\xi} \tag{13}$$

where $\mathbf{X}$ *is the* $N \times (q+1)$ *data matrix of all the predictors including the intercept.* $\mathbf{W}$ *is an* $N \times N$ *diagonal matrix with the jth diagonal element being*

$$\mathbf{W}_{jj} = \frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\varphi^2(\eta_j),$$

where $\eta_j = \alpha_0 + \sum_{k=1}^{q-1}\alpha_k x_{kj} + \alpha_q gEUD_j(n^*)$ *and* $\varphi(\cdot)$ *is the standard normal probability density function.* $\boldsymbol{\xi}$ *is a* $N \times 1$ *vector with the jth element* ξ_j *being:*

$$\xi_i = \frac{\eta_j\phi(\eta_j) - (w_1-1)\varphi(\eta_j)}{2\phi(\eta_j)}\mathbf{Q}_{jj},$$

where $\mathbf{Q}_{jj}$ *is the jth diagonal element of matrix* $\mathbf{Q} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$.

Using the result in Proposition 5, we can obtain an unbiased estimator for $\boldsymbol{\alpha}$, i.e., $\ddot{\boldsymbol{\alpha}} = \check{\boldsymbol{\alpha}} - \ddot{bias}(\check{\boldsymbol{\alpha}})$. $\ddot{bias}(\check{\boldsymbol{\alpha}})$ is an estimate for the theoretical bias in Equation (13) by using $\check{\eta}_j = \check{\alpha}_0 + \sum_{k=1}^{q-1}\check{\alpha}_k x_{kj} + \check{\alpha}_q gEUD_j(n^*)$. Furthermore, it can be shown that:

$$Var(\ddot{\boldsymbol{\alpha}}) = \left(\frac{N-q-1}{N}\right)^2 Var(\check{\boldsymbol{\alpha}}).$$

Given that the sample size N is typically larger than the number of parameters, $q+1$, we can get $Var(\ddot{\boldsymbol{\alpha}}) < Var(\check{\boldsymbol{\alpha}})$. This means that the variance of $\ddot{\boldsymbol{\alpha}}$ is smaller than $\check{\boldsymbol{\alpha}}$, i.e., the bias correction does not increase the variance of the estimator. Proof for the above variance relationship is similar to that for the case of generalized linear models (Cordeiro and McCullagh, 1991).

Finally, we summarize the steps of our proposed integrated model, as introduced in Sections 2.2 and 2.3, in Figure 3. The R code of our model has been submitted to GitHub website and is publicly available at https://github.com/xliu203/Integration-of-biological-model-and-statistical-model/blob/master/glmnet.probit_v2.R.

3. Case studies

3.1. Application to NTCP modeling of acute rectal complication for IMRT treatment of prostate cancer

We present an application in which patients were treated with IMRT for prostate cancer. A serious complication these patients may suffer from after the IMRT is the grade 2+ acute rectal complication with symptoms including anal pain, diarrhea, and rectal obstruction. We obtain a dataset of 86 patients from our collaborating institution, the Department of Radiation Oncology at the Mayo Clinic Arizona. The study was approved by the Institutional Review Board (IRB) of Mayo Clinic Arizona and included written informed consent from all subjects. All patients were diagnosed with prostate cancer. The IMRT they received was set up using the following protocol: A static field IMRT technique with seven coplanar 6 MV fields was employed. The whole prostate was designated as a Clinical Target Volume, and two Planning Target Volumes were created using uniform 3 mm and 6 mm expansions. A dose of 77.4 Gy in 43 fractions (1.8 Gy/fraction) was prescribed to the 3 mm expansion, and a dose of 70 Gy to the 6 mm expansion. Seminal Vesicles with uniform 7 mm expansion were prescribed 54 Gy. A Simultaneous Integrated Boost (SIB) was given to areas suspicious for cancer, as demonstrated in a planning multi-parametric MRI scan, which was a combination of T2-weighted imaging, diffusion weighted imaging and dynamic contrast-enhanced imaging. The SIB volume was identified by a diagnostic radiologist specializing in genitourinary imaging, was not expanded, and was prescribed a dose of 81–83 Gy. All patients were planned using the Eclipse Treatment Planning System (TPS) produced by Varian, Inc.

Table 1. Comparison between the AUCs of the integrated model, biological model (LKB), and statistical model (l_1-penalized logistic regression).

	Integrated model	*Biological model*	*Statistical model*
AUC on test data	0.82	0.66	0.76

Because we focus on the rectal complication, the rectum was drawn as a whole organ bounded by ischial tuberosity inferiorly and sigmoid flexure superiorly. Then, a DVH on the rectum was extracted for each patient using automated scripts that were written within the Applications Programmer Interface of the Eclipse TPS manufactured by Varian, Inc. Furthermore, we include 11 patient-specific variables that potentially affect the complication, which are age, concurrent treatment status, diabetes status, Gleason score, Androgen Deprivation Therapy (ADT) status, adjuvant ADT status, neoadjuvant ADT status, Prostate-Specific Antigen (PSA) level prior to treatment, prostate volume, use of statin medications, and stage of the disease (T-stage). After the IMRT, each patient's medical records were reviewed by a physician and his rectal complication was graded based on the Common Terminology Criteria for Adverse Events (CTCAE) v4.0. Among the 86 patients in the dataset, 23 developed the grade 2+ acute rectal complication.

We apply the integrated model to the dataset according to the steps in Figure 3. Recall that the integrated model includes a step for consistency correction, for the case where the population fraction of complication, τ, is smaller than the sample fraction of complication, $\bar{y}$. In our dataset, $\bar{y} = 23/86 = 26.7\%$, while $\tau < 20\%$ in the published literature on population-based studies (Tucker *et al.*, 2012). Therefore, consistency correction is needed. Furthermore, to choose the optimal tuning parameters, we adopt a model selection criterion called AICc, which includes a correction for the original AIC under small sample sizes (Hurvich and Tsai, 1989). The results from the integrated model are as follows: Among all the patient-specific variables included in the dataset, six are selected using AICc: diabetes status, prostate volume, PSA, statins use, ADT status, T-stage. The optimal n^* is found to be 0.154. These results are consistent with findings in the literature. For example, statins are a class of drugs often prescribed by doctors to help lower cholesterol levels in the blood. Statins use is negatively related to the probability of complication, indicating that the use of statins might be protective against the development

of the 2+ acute rectal complication by patients. At least one biological mechanism behind this seemingly protective effect has been suggested (Malek, 2015), and a relatively recent study reported a similar result, namely a negative association between acute rectal complication during pelvic RT and the use of statins (Wedlake *et al.*, 2012). This corroborates our finding. PSA is a blood test that is commonly used to detect prostate cancer; the higher the level of PSA, the higher the chance the patient has prostate cancer. Finally, knowing that the range of n is between zero and one, the optimal $n^* = 0.154$ found by the integrated model is small. This is consistent with prior findings (Bentzen *et al.*, 2010) and agrees with clinical expectation; the rectum is a serial organ, and it is well-known that serial organs tend to have small n (Gulliford *et al.*, 2012; Li *et al.*, 2012).

Furthermore, we would like to assess the prediction/classification accuracy of the integrated model in comparison with the biological and statistical models used alone. The best strategy in accuracy assessment of a statistical model without running the risk of overfitting is to divide the entire dataset into a training set and a test set. Samples in the test set are not used in training, but rather are only used to compute the classification accuracy of the training model. Realizing that we have a small dataset, we put all but one samples in the training set and the remaining sample in the test set. This will allow us to compute the classification accuracy on one test sample. We repeat this training–test split over all the samples, which will allow us to compute the test accuracy on all samples. Note that this scheme is different from leave-one-out cross-validation, as the latter would report the best accuracy optimized on the test set whereas our scheme assumes the test set is completely unseen at the training stage. Using this scheme, Table 1 shows the test accuracy of the integrated model in comparison with those of the biological and statistical models. The accuracy metric is Area Under the Curve (AUC). The use of AUC avoids having to choose a cutoff for the predicted probabilities, and therefore, provides a more objective measure for the prediction accuracy. Recall that LKB model fitting only uses the DVH and cannot take patient-specific variables into consideration. As a result, the AUC is low. The commonly used statistical model in the NTCP literature is logistic regression (Boomsma *et al.*, 2012; Deville *et al.*, 2012; Cella *et al.*, 2013). We follow the convention of NTCP modeling and build a logistic regression that includes $D_{10\%}$, $D_{15\%}$, $D_{20\%}$, $\cdots$, $D_{90\%}$ and patient-specific variables as predictors. Here, $D_{\alpha\%}$ is the radiation dose such that $\alpha\%$ of the rectal volume receives this dose level or higher. For a fair comparison with the integrated model, we also use an l_1-penalty for variable selection and select the penalty parameter using AICc. The AUC of the statistical model is around 0.76, which is also lower than the integrated model.

Finally, we report the AUCs of the integrated model under other model selection criteria. The AUCs under AIC and BIC are 0.82 and 0.78, respectively, which are still higher than the biological and statistical models used alone. The AUC under AIC is the same as that under AICc (Table 1), implying that the integrated model is not sensitive to the small sample correction that is accounted for by AICc. The AUC under BIC is relatively lower than other criteria. BIC is known to be the most suitable option for cases where there is a very large number of predictors, whereas our study involves only 11 predictors.

3.2. Simulation experiments

We would like to compare the performance of the integrated model with biological and statistical models under different parameter settings using simulation data. A significant challenge of this study is how to simulate DVH for each virtual patient. To make sure that the simulated DVH has similar data characteristics to the DVH of real patients, we simulate the DVH of each virtual patient from the DVH measurements of the real patients. Specifically, we fit a multivariate normal distribution for a random vector of $(D_{1\%}, D_{2\%}, \ldots, D_{99\%})^T$ using the DVH of the real patients. Next, we sample from the fitted distribution to create the DVH for a virtual patient. Then, the DVH is used to compute $gEUD(n)$ using Equation (1). We set $n = 0.2$ and 0.3 in our simulation experiments. Furthermore, we sample from a multivariate normal distribution, $(X_1, \ldots, X_p)^T \sim N_p(0, \mathbf{\Sigma})$, to create data for p patient-specific variables of the virtual patient. $\mathbf{\Sigma}$ has all its diagonal elements being one and off-diagonal elements being $\mathbf{\Sigma}_{ij} = 0.2^{|i-j|}$ to account for possible correlation between the variables. We set $p = 14$ in our experiment, which is close to the number of patient-specific variables included in the real-data application in Section 3.1.

Furthermore, to set the coefficients for the patient-specific variables and $gEUD(n)$, we refer to the model in Equation (14) for appropriate ranges of the coefficients. Ten out of the 14 patient-specific variables are set to have zero coefficients. In this way, we can test the accuracy of the model in selecting patient-specific variables. In the remaining four patient-specific variables with non-zero coefficients, two are set to have positive coefficients and the other two are set to have negative coefficients. The magnitude of non-zero coefficients is set to be $\alpha_k = 0.8,\ k = 1, 2, 3, 4$. We also run experiments on a smaller magnitude of 0.6. The coefficient for $gEUD(n)$ is set to be $\alpha_{p+1} =$ 0.25 and 0.3. The intercept α_0 is set to achieve a desired population fraction of complication, τ. A value of $\tau = 20\%$ is used in our experiments.

With simulated data for a virtual patient and under a particular setting of the model coefficients, we use the right-hand side of Equation (6) to generate the probability that the patient develops the complication. Then, this probability is used as the parameter of a Bernoulli distribution, from which a binary variable y can be sampled. Furthermore, to mimic the reality that the sample fraction of complication in a dataset, $\bar{y}$, is usually higher than the population fraction of complication, τ, we set $\bar{y} = 40\%$. The sample size of the dataset is 200.

We apply the integrated model, LKB, and l_1-penalized logistic regression to the simulation datasets. The results are

Table 2. Comparison between the integrated model, biological model, and statistical model ($\alpha_k = 0.8,\ \alpha_{p+1} = 0.25,\ n = 0.2$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment. $p < 0.001^*$ is the p value for a one-sided hypothesis testing that the integrated model has a higher AUC than the biological or statistical model whichever has a higher AUC.

	LOOCV-AUC	*Sensitivity in selecting patient-specific variables*	*Specificity in selecting patient-specific variables*	*95% Confidence interval for n*
Integrated model	0.91 (0.009) $p < 0.001^*$	1 (0.000)	0.85 (0.146)	$0.2 \in [0.078, 0.368]$
Biological model	0.68 (0.005)	—	—	$0.2 \notin [0.060, 0.184]$
Statistical model	0.86 (0.016)	1 (0.000)	0.84 (0.142)	—

Table 3. Comparison between the integrated model, biological model, and statistical model ($\alpha_k = 0.6,\ \alpha_{p+1} = 0.25,\ n = 0.2$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment. $p < 0.001^*$ is the p value for a one-sided hypothesis testing that the integrated model has a higher AUC than the biological or statistical model whichever has a higher AUC.

	LOOCV-AUC	*Sensitivity in selecting patient-specific variables*	*Specificity in selecting patient-specific variables*	*95% Confidence interval for n*
Integrated model	0.88 (0.029) $p < 0.001^*$	0.94 (0.210)	0.87 (0.121)	$0.3 \in [0.211, 0.389]$
Biological model	0.74 (0.001)	—	—	$0.3 \notin [0.166, 0.178]$
Statistical model	0.77 (0.052)	0.94 (0.190)	0.86 (0.126)	—

Table 4. Comparison between the integrated model, biological model, and statistical model ($\alpha_k = 0.8,\ \alpha_{p+1} = 0.3,\ n = 0.3$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment. $p < 0.001^*$ is the p value for a one-sided hypothesis testing that the integrated model has a higher AUC than the biological or statistical model whichever has a higher AUC.

	LOOCV-AUC	*Sensitivity in selecting patient-specific variables*	*Specificity in selecting patient-specific variables*	*95% Confidence interval for n*
Integrated model	0.93 (0.009) $p < 0.001^*$	1 (0.000)	0.83 (0.141)	$0.3 \in [0.131, 0.474]$
Biological model	0.75 (0.001)	—	—	$0.3 \notin [0.166, 0.184]$
Statistical model	0.84 (0.030)	0.98 (0.106)	0.85 (0.099)	—

Table 5. Comparison between the integrated model, biological model, and statistical model ($\alpha_k = 0.6,\ \alpha_{p+1} = 0.3,\ n = 0.3$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment. $p < 0.001^*$ is the p value for a one-sided hypothesis testing that the integrated model has a higher AUC than the biological or statistical model whichever has a higher AUC.

	LOOCV-AUC	*Sensitivity in selecting patient-specific variables*	*Specificity in selecting patient-specific variables*	*95% Confidence interval for n*
Integrated model	0.89 (0.015) p $< 0.001^*$	0.98 (0.076)	0.86 (0.134)	$0.2 \in [0.105, 0.342]$
Biological model	0.71 (0.003)	—	—	$0.2 \notin [0.114, 0.176]$
Statistical model	0.81 (0.031)	0.96 (0.105)	0.86 (0.152)	—

shown in Tables 2 to 5. The following observations can be drawn:

First, the AUC of the integrated model is significantly higher than LKB and logistic regression (p value <0.001) across all the simulation settings. The AUC of the integrated model becomes lower when the magnitude of the coefficients for patient-specific variables, α_k, gets smaller, but it is little affected by n and the coefficient for $gEUD(n)$, α_{p+1}. Moreover, although LKB has the lowest mean AUC among the three models, the standard deviation of the AUC over 50 repetitions of the simulation experiment for LKB is the smallest. This is expected because LKB, as a biological model, is built upon radiobiological principles such that its performance is less affected by sampling variability. On the other hand, logistic regression has the largest standard deviation of the AUC. This is because statistical models are purely data-driven, and therefore, the performance is more variable across different datasets. By integrating the biological and statistical models, the proposed integrated model can achieve a high accuracy, due to the inclusion of patient-specific variables, and a more robust performance against sampling variability, due to the consideration of radiobiological principles.

Furthermore, the integrated model achieves high sensitivity and specificity in selecting the patient-specific variables across all the experiments. Here, sensitivity is the proportion of non-zero coefficients for patient-specific variables that are correctly identified as being non-zero. Specificity is the proportion of zero coefficients for patient-specific variables that are correctly identified as being zero. The sensitivity and specificity become lower when the magnitude of the coefficients for patient-specific variables, α_k, gets smaller. The same phenomenon is observed if we keep α_k unchanged, but decrease the sample size. These observations are consistent with findings on existing variable selection approaches (Huang *et al.*, 2012). Logistic regression achieves similar levels of sensitivity and specificity to the integrated model. Considering this result, together with that on AUC, we can conclude that logistic regression may perform as well as the integrated model on a single dataset, which is reflected by the sensitivity and specificity in selecting patient-specific variables. However, it performs worse than the integrated model when one wants to apply the model trained on one dataset to another dataset (i.e., weaker generalizability), which is reflected by the AUC.

Finally, we compare the estimated parameter n between the integrated model and LKB. A universally true observation across all the experiments is that the confidence interval of n includes the true value in the integrated model, but not in LKB. As LKB fails to account for the effect of patient-

specific variables on the probability of complication, its parameter estimation is compromised. This result corroborates the low AUC of LKB.

4. Conclusion

In this article, we proposed an integrated model for NTCP modeling. We developed the model by starting with an in-depth understanding of the biological model (i.e., LKB) parameters. Among all the parameters, TD_{50} reflects the radiation dose a patient can tolerate before developing complication with a chance higher than a random guess, and therefore should be patient-specific. We proposed to link patient-specific variables with TD_{50} by a linear model and used this personalized TD_{50} to replace the original TD_{50} in LKB. This resulted in an integrated model formulation. We further added to the formulation a sparsity-inducing penalty to enable variable selection from high-dimensional patient-specific variables, and a biological constraint on the model coefficients to account for the fact that radiation dose always poses at least "some" risk of complication to normal tissue. Next, we developed an efficient algorithm to estimate the parameters of the integrated model. Furthermore, we performed theoretical analysis and proposed modified approaches to ensure that the integrated model had statistically consistent and unbiased coefficient estimators. Finally, we applied the integrated model to a real dataset of prostate cancer patients treated with IMRT who are at risk of developing the grade 2+ acute rectal complication. The integrated model had higher prediction accuracy measured by AUC on test data than the biological (i.e., LKB) and statistical models (i.e., l_1-penalized logistic regression) used in isolation. Various simulation studies were also conducted, showing that the integrated model significantly outperformed both biological and statistical models. The variable selection sensitivity and specificity of the integrated and statistical models were comparable. These results indicated that the statistical model may perform as well as the integrated model on a single dataset, but it has worse generalizability to other studies. In addition, the integrated model accurately estimated the organ parameter, whereas LKB was not able to do so.

There are some limitations of the present study, which drive future investigation. The proposed model was demonstrated on a small dataset consisting of 86 patients with 23 individuals having complications. More data are needed to further validate the model and findings. Also, our study found PSA to be a significant predictor for rectal complications in prostate cancer patients. We are not aware of any prior study that reported this relationship, although there are plenty of studies that correlate PSA with the existence of prostate cancer (Partin *et al.*, 1996; Catalona *et al.*, 1997, 2000). Further investigation is needed to validate this finding with more data and discover the biological mechanism behind this relationship if it still holds true.

Funding

Funding for this research was provided by the Mayo Clinic and the NSF under grant 1149602.

Notes on contributors

Xiaonan Liu is a Ph.D. student in industrial engineering at Arizona State University. He received his B.S. from the University of Science and Technology of China in 2013 and M.S. in industrial engineering from Arizona State University in 2015. His research focuses on statistical modeling and machine learning with applications in heathcare and semiconductor manufacturing. He is a student member of IISE, INFORMS, and IEEE.

Mirek Fatyga is an associate professor in radiation oncology at the Mayo Clinic Arizona. He received his M.S. in physics from the University of Warsaw, Poland and his Ph.D. in nuclear physics from Indiana University. His current research interests include applications of statistical analysis of clinical outcomes in radiation therapy to individualized treatments of cancer patients. He is also serving as a Chair of Physics Division in the Department of Radiation Oncology, Mayo Clinic Arizona. He is a member of the American Association of Physicists in Medicine.

Teresa Wu is a professor in industrial engineering at Arizona State University. She received her Ph.D. in industrial engineering from the University of Iowa in 2001. Her current research interests include swarm intelligence, distributed decision support, and health informatics. She is currently serving as the editor-in-chief for *IISE Transactions on Healthcare Systems Engineering.*

Jing Li is an associate professor in industrial engineering at Arizona State University. She received her B.S. from Tsinghua University in China and an M.A. in statistics and a Ph.D. in industrial and operations engineering from the University of Michigan in 2005 and 2007, respectively. Her research interests are statistical modeling and machine learning for health care applications. She is a recipient of an NSF CAREER award. She is a member of IISE, INFORMS, and IEEE.

References

Bentzen, S.M., Constine, L.S., Deasy, J.O., Eisbruch, A., Jackson, A., Marks, L.B., … Yorke, E.D. (2010) Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *International Journal of Radiation Oncology• Biology• Physics*, **76**(3), S3–S9.

Boomsma, M.J., Bijl, H.P., Christianen, M.E., Beetz, I., Chouvalova, O., Steenbakkers, R.J., … Langendijk, J.A. (2012) A prospective cohort study on radiation-induced hypothyroidism: Development of an NTCP model. *International Journal of Radiation Oncology• Biology• Physics*, **84**(3), e351–e356.

Catalona, W.J., Southwick, P.C., Slawin, K.M., Partin, A.W., Brawer, M.K., Flanigan, R.C. … Bray, K.R. (2000) Comparison of percent free PSA, PSA density, and age-specific PSA cutoffs for prostate cancer detection and staging. *Urology*, **56**(2), 255–260.

Catalona, W.J., Smith, D.S. and Ornstein, D.K. (1997) Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: Enhancement of specificity with free PSA measurements. *Jama*, **277**(18), 1452–1455.

Cella, L.M, D'Avino, V., Liuzzi, R., Conson, M., Doria, F., Faiella, A., … Pacelli, R. (2013) Multivariate normal tissue complication probability modeling of gastrointestinal toxicity after external beam radiotherapy for localized prostate cancer. *Radiation Oncology*, **8**(1), 221.

Cordeiro, G.M. and McCullagh, P. (1991) Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(3), 629–643.

Deasy, J.O. (2000) Comments on the use of the Lyman-Kutcher-Burman model to describe tissue response to nonuniform irradiation. *International Journal of Radiation Oncology• Biology• Physics*, **47**(5), 1458–1459.

Deville, C., Both, S., Bui, V., Hwang, W.T., Tan, K.S., Schaer, M., Tochner, Z. and Vapiwala, N. (2012) Acute gastrointestinal and genitourinary toxicity of image-guided intensity modulated radiation therapy for prostate cancer using a daily water-filled endorectal balloon. *Radiation Oncology*, **7**(1), 76.

Gulliford, S.L., Partridge, M., Sydes, M.R., Webb, S., Evans, P.M. and Dearnaley, DP. (2012) Parameters for the Lyman Kutcher Burman (LKB) model of normal tissue complication probability (NTCP) for specific rectal complications observed in clinical practise. *Radiotherapy and Oncology*, **102**(3), 347–351.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015) *Statistical Learning with Sparsity: The LASSO and Generalizations*, Chapman and Hall/CRC, Boca Raton, FL.

Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, **27**(4), 481–499.

Hurvich, C.M. and Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297–307.

Källman, P., Ågren, A. and Brahme, A. (1992) Tumour and normal tissue responses to fractionated non-uniform dose delivery. *International Journal of Radiation Biology*, **62**(2), 249–262.

King, G. and Zeng, L. (2001) Logistic regression in rare events data. *Political Analysis*, **9**(2), 137–163.

Li, X.A., Alber, M., Deasy, J.O., Jackson, A., Ken Jee, K.W., Marks, L.B., … Yorke, E.D. (2012) 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. *Medical Physics*, **39**(3), 1386–1409.

Liu, M., Moiseenko, V., Agranovich, A., Karvat, A., Kwan, W., Saleh, Z.H., Apte, A.A. and Deasy, J.O. (2010) Normal tissue complication probability (NTCP) modeling of late rectal bleeding following external beam radiotherapy for prostate cancer: A test of the QUANTEC-recommended NTCP model. *Acta Oncologica*, **49**(7), 1040–1044.

Lyman, J.T. (1985) Complication probability as assessed from dose-volume histograms. *Radiation Research*, **104**(2), 13–19.

Malek, K. (2015) Contribution to the clinical and laboratory study of tissue reactions observed radiosensitivity after radiotherapy of prostate cancer: Potential radioprotective effect of statins. Doctoral Thesis, University Joseph Fourier – Grenoble, Grenoble, France.

Manski, C.F. and Lerman, S.R. (1977) The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, **45**(8), 1977–1988.

Michalski, J.M., Gay, H., Jackson, A., Tucker, S.L. and Deasy, J.O. (2010) Radiation dose–volume effects in radiation-induced rectal injury. *International Journal of Radiation Oncology• Biology• Physics*, **76**(3), S123–S129.

Partin, A.W., Catalona, W.J., Southwick, P.C., Subong, E.N., Gasior, G.H. and Chan, D.W. (1996) Analysis of percent free prostate-specific antigen (PSA) for prostate cancer detection: Influence of total PSA, prostate volume, and age. *Urology*, **48**(6), 55–61.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tucker, S.L., Liu, H. H., Liao, Z., Wei, X., Wang, S., Jin, H., Komaki, R., Martel, M., Mohan, R. (2008) Analysis of radiation pneumonitis risk using a generalized Lyman model. *International Journal of Radiation Oncology• Biology• Physics*, **72**(2), 568–574.

Tucker, S.L Michalski, J. M., Bosch, W. R., Mohan, R., Dong, L., Winter, K., Purdy, J.A., Cox, J. D. (2012) Use of fractional dose–volume histograms to model risk of acute rectal toxicity among patients treated on RTOG 94-06. *Radiotherapy and Oncology*, **104**(1), 109–113.

Wedlake, L., Silia, F., Benton, B., Lalji, A., Thomas, A., Dearnaley, D.P., Blake, P., Tait, D., Khoo, V.S., Andreyev, H.J.N. (2012) Evaluating the efficacy of statins and ACE-inhibitors in reducing gastrointestinal toxicity in patients receiving radiotherapy for pelvic malignancies. *European Journal of Cancer*, **48**(14), 2117–2124.

Appendices

A: Proof of Proposition 1

Let $a = 1/n$ and $D = \max_i D_i$. By definition, $D_i \le D$ for $\forall i$. Therefore, $\left(\sum_i v_i D_i^a\right)^{1/a} \le \left(\sum_i v_i D^a\right)^{1/a} = D$, i that is

$$\limsup_{a\to\infty} \left(\sum\nolimits_i v_i D_i^a\right)^{1/a} \le D. \tag{A1}$$

Furthermore, let $S_\delta = \{i | D \le D_i + \delta\}$. Then, $\left(\sum_i v_i D_i^a\right)^{1/a} \ge \left(\sum_{i\in S_\delta} v_i\right)^{1/a}(D-\delta)$. By letting $\delta \to 0$ and $a \to \infty$, we can get

$$\liminf_{a\to\infty} \left(\sum\nolimits_i v_i D_i^a\right)^{1/a} \ge D. \tag{A2}$$

Combining Equations (A1) and (A2), we get $\lim_{a\to\infty}\left(\sum_i v_i D_i^a\right)^{1/a} = D$.

B: Proof of Proposition 2

If $\alpha^*_{p+1} \ge 0$, the constraint in the optimization problem in Equation (9) is automatically satisfied by the optimal solution to the unconstrained optimization problem. This means that the optimal solution to the unconstrained problem is just that to the constrained problem, i.e., $(\hat\alpha_0, ..., \hat\alpha_{p+1}) = (\alpha^*_0, ..., \alpha^*_{p+1})$.

If $\alpha^*_{p+1} < 0$, let $\boldsymbol{\alpha}^* = (\alpha^*_0, \cdots, \alpha^*_{p+1})$ and $\hat{\boldsymbol{\alpha}} = (\hat\alpha_0, \cdots, \hat\alpha_{p+1})$ be the optimal solutions to the unconstrained and constrained optimization problems, respectively. We need to prove that $(\hat\alpha_0, ..., \hat\alpha_p) = (\alpha^*_0, ..., \alpha^*_p)$ and $\hat\alpha_{p+1} = 0$. To prove this, we start by constructing a feasible solution to the constrained optimization in Equation (9), that is

$$\tilde{\boldsymbol{\alpha}} = \beta\hat{\boldsymbol{\alpha}} + (1-\beta)\boldsymbol{\alpha}^*, \tag{A3}$$

where

$$\beta = -\frac{\alpha^*_{p+1}}{\hat\alpha_{p+1} - \alpha^*_{p+1}}.$$

Here, β is valid because we know that $\alpha^*_{p+1} < 0$ and $\hat\alpha_{p+1} \ge 0$ such that the denominator of β must not be zero. Also, $\tilde{\boldsymbol{\alpha}}$ is a feasible solution to Equation (9) because:

$$\tilde\alpha_{p+1} = \beta\hat\alpha_{p+1} + (1-\beta)\alpha^*_{p+1} = 0, \tag{A4}$$

which satisfies the constraint in Equation (9). Then, the following inequality holds:

$$g(\tilde{\boldsymbol{\alpha}}) \ge g(\hat{\boldsymbol{\alpha}}) \ge g(\boldsymbol{\alpha}^*). \tag{A5}$$

The first "$\ge$" in Equation (A5) holds because $\tilde{\boldsymbol{\alpha}}$ is a feasible solution and $\hat{\boldsymbol{\alpha}}$ is the optimal solution. The second "$\ge$" holds because $\hat{\boldsymbol{\alpha}}$ is the optimal solution to the constrained optimization while $\boldsymbol{\alpha}^*$ is that to the unconstrained optimization.

Furthermore, let $g(\alpha_0, \cdots, \alpha_{p+1}) = -l(\alpha_0, \cdots, \alpha_{p+1}) + \lambda\sum_{k=1}^p |\alpha_k|$, which is the objective function of Equation (9). $g(\cdot)$ a convex function. Therefore, we can get

$$\begin{aligned} g(\tilde{\boldsymbol{\alpha}}) &= g(\beta\hat{\boldsymbol{\alpha}} + (1-\beta)\boldsymbol{\alpha}^*) \\ &\le \beta g(\hat{\boldsymbol{\alpha}}) + (1-\beta)g(\boldsymbol{\alpha}^*) \\ &\le \beta g(\hat{\boldsymbol{\alpha}}) + (1-\beta)g(\hat{\boldsymbol{\alpha}}) = g(\hat{\boldsymbol{\alpha}}). \end{aligned} \tag{A6}$$

The fact that $g(\tilde{\boldsymbol{\alpha}}) \ge g(\hat{\boldsymbol{\alpha}})$ in Equation (A5) and $g(\tilde{\boldsymbol{\alpha}}) \le g(\hat{\boldsymbol{\alpha}})$ in Equation (A6) leads to $g(\tilde{\boldsymbol{\alpha}}) = g(\hat{\boldsymbol{\alpha}})$. This means that $\tilde{\boldsymbol{\alpha}}$ is the optimal

solution to Equation (9), i.e., $\hat{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}}$. Then, $\hat{\alpha}_{p+1} = \widetilde{\alpha}_{p+1} = 0$ according to Equation (A4). Furthermore, using Equation (A3), we can get $(\hat{\alpha}_0, ..., \hat{\alpha}_p) = (\alpha_0^*, ..., \alpha_p^*)$. □

C: Proof of Proposition 5

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for the linear coefficients $\boldsymbol{\beta}$ in a Generalized Linear Model (GLM). It can be derived that under a finite sample size, the bias of $\hat{\boldsymbol{\beta}}$ is

$$bias(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\xi}. \tag{A7}$$

Here,

$$\mathbf{X^T}\mathbf{W}\mathbf{X} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right],$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function. The jth element of $\boldsymbol{\xi}$ is

$$\xi_j = -\frac{1}{2}\frac{\mu_j''}{\mu_j'}\mathbf{Q}_{jj},$$

where $\mu_j' = \partial\mu_j/\partial\eta_j$, $\mu_j'' = \partial^2\mu_j/\partial\eta_j^2$, μ_j is the mean of the distribution from the exponential family the GLM corresponds to, η_j is the linear predictor of the GLM, and $\mathbf{Q}_{jj}$ is the jth diagonal element of matrix $\mathbf{Q} = \mathbf{X}(\mathbf{X^T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X^T}$. Next, we use the result in Equation (A7) to derive the bias of our model, i.e., $bias(\breve{\boldsymbol{\alpha}})$.

Specifically, the log-likelihood function of our model, $l_{n^*}^w(\boldsymbol{\alpha})$, is given in Equation (12). The first derivative of $l_{n^*}^w(\boldsymbol{\alpha})$ is

$$\frac{\partial l_{n^*}^w(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = \sum_{j=1}^N \left\{\frac{w_1 y_j}{\phi(\eta_j)} - \frac{w_0(1-y_j)}{1-\phi(\eta_j)}\right\}\varphi(\eta_j)\mathbf{x}_j$$

$$= \sum_{j=1}^N \left\{\frac{\left[w_1 y_j - w_0\phi(\eta_j)\right] - (w_1 - w_0)y_j\phi(\eta_j)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\right\}\varphi(\eta_j)\mathbf{x}_j,$$

where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. Furthermore, we can get

$$\mathbf{X^T}\mathbf{W}\mathbf{X} = -E\left[\frac{\partial^2 l_{n^*}^w(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right] = \sum_{j=1}^N \left\{\frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\right\}\varphi^2(\eta_j)\mathbf{x}_j\mathbf{x}_j^T. \tag{A8}$$

It is clear from Equation (A8) that $\mathbf{W}$ is a diagonal matrix with the jth element being:

$$\mathbf{W}_{jj} = \frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\varphi^2(\eta_j). \tag{A9}$$

Next, to derive ξ_i, we need to derive μ_i' and μ_i'' for our model. Specifically, $\mu_i = \phi(\eta_i)^{w_1}$. Therefore, the first derivative of μ_i is

$$\mu_i' = w_1\phi(\eta_i)^{w_1-1}\varphi(\eta_i),$$

and the second derivative of μ_i is

$$\mu_i'' = -w_1\phi(\eta_i)^{w_1-1}\varphi(\eta_i)\eta_i + w_1(w_1-1)\phi(\eta_i)^{w_1-2}\varphi^2(\eta_i).$$

Therefore,

$$\xi_j = -\frac{1}{2}\frac{\mu_j''}{\mu_j'}\mathbf{Q}_{jj} = \frac{\eta_j\phi(\eta_j) - (w_1-1)\varphi(\eta_j)}{2\phi(\eta_j)}\mathbf{Q}_{jj}. \tag{A10}$$

Finally, by inserting Equations (A9) and (A10) into Equation (A7), we can obtain the $bias(\breve{\boldsymbol{\alpha}})$ in Equation (13). □