Radiotherapy and Oncology 123 (2017) 85–92

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Radiation induced lung injury

# Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis

Yi Luo [a], Issam El Naqa [a], Daniel L. McShan [a], Dipankar Ray [a], Ines Lohse [a], Martha M. Matuszak [a], Dawn Owen [a], Shruti Jolly [a], Theodore S. Lawrence [a], Feng-Ming (Spring) Kong [b], Randall K. Ten Haken [a,*]

[a] Department of Radiation Oncology, The University of Michigan, Ann Arbor; and [b] Department of Radiation Oncology, Indiana University, Indianapolis, United States

ARTICLE INFO

*Article history:*
Received 22 July 2016
Received in revised form 5 January 2017
Accepted 5 February 2017
Available online 22 February 2017

*Keywords:*
Lung cancer
Radiation pneumonitis
Bayesian network analysis
Biophysical interactions

ABSTRACT

*Background:* In non-small-cell lung cancer radiotherapy, radiation pneumonitis ≥ grade 2 (RP2) depends on patients' dosimetric, clinical, biological and genomic characteristics.
*Methods:* We developed a Bayesian network (BN) approach to explore its potential for interpreting biophysical signaling pathways influencing RP2 from a heterogeneous dataset including single nucleotide polymorphisms, micro RNAs, cytokines, clinical data, and radiation treatment plans before and during the course of radiotherapy. Model building utilized 79 patients (21 with RP2) with complete data, and model testing used 50 additional patients with incomplete data. A developed large-scale Markov blanket approach selected relevant predictors. Resampling by *k*-fold cross-validation determined the optimal BN structure. Area under the receiver-operating characteristics curve (AUC) measured performance.
*Results:* Pre- and during-treatment BNs identified biophysical signaling pathways from the patients' relevant variables to RP2 risk. Internal cross-validation for the pre-BN yielded an AUC = 0.82 which improved to 0.87 by incorporating during treatment changes. In the testing dataset, the pre- and during AUCs were 0.78 and 0.82, respectively.
*Conclusions:* Our developed BN approach successfully handled a high number of heterogeneous variables in a small dataset, demonstrating potential for unraveling relevant biophysical features that could enhance prediction of RP2, although the current observations would require further independent validation.



Identification of biophysical or pan-Omics models combining clinical, physical, molecular cancer profiling and computational analysis should enable "personalized medicine," where treatment strategies can be individually tailored based on combinations of diverse existing data resources[1,2]. However, large scale and diverse radiobiological datasets need to be integrated within a biophysical signaling network of response, interactive manipulation of such network should be supported, and mechanistic radiobiological insights from the analysis would be required. Dose limiting toxicities in the form of radiation pneumonitis (RP) and pulmonary fibrosis are major obstacles in the successful radiation treatment of lung neoplasms. Radiosensitivity of the lung has been shown to depend on multiple factors that include patient's clinical, physical, imaging and molecular characteristics [3–10]. The availability of clinical data, laboratory information, and diagnostic images before and during the course of radiotherapy provides opportunities as well as challenges to build a predictive model for patient specific risks of RP. The purpose of this paper is to develop a systematic methodology based on graphical analysis by Bayesian networks to explore its potential for unraveling biophysical signaling interactions of radiation pneumonitis (RP) from a heterogeneous dataset.

Bayesian networks (BNs) are probabilistic graphical models that represent a set of biophysical variables and their conditional dependencies via a directed acyclic graph (DAG) [11]. BNs are also called belief networks, where, in a graphical representation, each node represents a random variable and the edges between the nodes represent probabilistic dependences among the interacting random variables, analogous to using nomograms for representing linear/logistic models in traditional statistical analysis. Thus, BNs are able to explore probabilistic relationships among multiple interacting variables [12]. However, due to limited sample sizes together with large numbers of parameters that need to be optimized in a typical clinical application, an inherent limitation in BN analyses is the need to transform continuous variables into discrete ones, which may lead to loss of information. To mitigate this

* Corresponding author at: University of Michigan, Department of Radiation Oncology, UH B2C432, SPC 5010, 1500 East Medical Center Dr., Ann Arbor, MI 48109-5010, United States.
*E-mail address:* rth@med.umich.edu (R.K. Ten Haken).

http://dx.doi.org/10.1016/j.radonc.2017.02.004
0167-8140/

effect, Hartemink's algorithm, which employs a mutual information approach to retain the most permissible information, can be used for discretization of the data, as it was here. In all, as BNs can effectively probe the interactions of clinical, physical, and molecular data, accommodate hierarchal relationships embedded in radiation response signaling pathways of RP development, and handle missing data, they are promising approaches to support radiation treatment decision making. Previously, a BN was employed to model response in lung cancer including RP [13]. However, that study was limited in scope to 16 selected features that included cytokines and dosimetric variables only. The current study has 200 features, also including genetic single nucleotide polymorphisms (SNPs) and micro RNA (miRNA) markers. Due to the high dimensional nature of this retrospective dataset, different developed methods for feature selection (extended Markov blanket) and BN structure learning (Tabu Search) are used to handle large-scale analysis. In addition to internal re-sampling cross-validation using the training data, the current study includes a testing phase, which also assesses the use of the BN for marginalization of missing information.

## Materials and methods

### *Patient samples*

This study involved secondary analysis of NSCLC patients that had been treated on four prospective protocols under IRB approval. The second and third treated patients to standard doses (<74 Gy at 2 Gy per fraction) and the first and fourth were dose escalation studies (total doses up to of 86 Gy over 30 fractions). Radiation dose distributions for total uninvolved lung (left plus right exclusive of gross tumor volume) were computed within the Varian Eclipse treatment planning system using the AAA dose algorithm. All total dose values were converted to their 2 Gy equivalents (EQD2) by the linear-quadric model using locally developed software. Dosimetric features with different parameters were highly correlated with each other in the MB in early screening, and obscured the selection approach to identify important features from other categories for RP2 prediction. Thus, Mean_Lung_Dose computed from EQD2 dose distributions generated with an $\alpha/\beta$ of 4 Gy was considered to be representative of the dosimetric features [14].

Blood samples were obtained at baseline and after approximately 1/3 and 2/3 of the scheduled radiation doses were completed. Pre-treatment blood samples were analyzed for cytokine levels [15], miRNAs [16], and SNPs [17,18], which have been identified as candidates from the literature as related to lung cancer or inflammatory disease (Table 1 in Appendix A). The slopes of cytokine changes from before to during treatment were also determined as patients' responses to radiation treatment.

The patients' RP were scored by five grades (CTCAE 3.0) based on clinical assessment and imaging findings [19], and here, RP with grade 2 or above (RP2) was used to represent complication of radiation treatment. The study includes 79 NSCLC patients (with 21 cases of RP2) with complete data (200 features) utilized for biophysical pathway model building and 50 additional patients (with 3 cases of RP2) with incomplete data (missing some blood sample analyses) used for model testing. Input data pre-processing is detailed in Appendix B.

### *Development of BN approach to identify appropriate biophysical interactions*

#### *Step 1. Selection of relevant variables for RP modeling*

For feature selection, we used a modified version of the Markov blanket (MB) approach which represents not only the inner family of inter-related variables as commonly practiced [11] but also their extended family, i.e., next-of-kin variable relationships. The MB of RP2 represents the features most related to RP2, and RP2 is independent from the rest of other features given its MB [20]. However, only including the mostly related features for RP2 prediction may result in loss of extended relationships of higher orders for prediction purposes. Therefore, to minimize this effect, we also extended the MB by one order (next-of-kin MB) for each feature in the original MB. In addition to finding good feature subsets, the MB could also help determine causal relationships among various nodes in a BN [21]. Here, the MBs of RP2 and next-of-kin were found using the HITON algorithm, which is a fast forward selection technique for neighborhood detection designed to exclude irrelevant nodes efficiently based on marginal associations [22]. This algorithm is particularly useful when the number of features is large compared to the number of samples as in the current case.

#### *Step 2. Identifying appropriate biophysical interactions for RP2 prediction*

A BN was employed to integrate the potential variables for RP2 prediction based on the existing data, which is a complex process given the limited sample size. A Tabu Search is a metaheuristic search method employing local search methods used for mathematical optimization [23], and was used in our study to generate a stable BN structure to integrate the selected biophysical variables from bootstrap samples of the original data. Although there are many biophysical interactions related to RP2, and the relationships among potential variables on each interaction are complicated, there existed many known prior biophysical relationships related to RP2 onset that could be exploited to constrain the search space in building the BN structure. These pathways or relationships may have their specific variables from different categories. The known causal influence relationships between them were considered as "biophysical rules" in our BN structure identification, which could be obtained from the literature. An example to identify appropriate biophysical interactions for RP2 prediction is detailed in Appendix C. Therefore, for a selected set of biophysical variables related to RP2, radiobiologically plausible relationships were built among them by adding constraints to the Tabu Search and forcing them to follow the biophysical rules.

Here BN structure learning was intended to find appropriate biophysical relationships to improve RP2 prediction performance by balancing variance reduction and loss of information. While each predictor in the stable BN can contribute to biophysical information for RP2 prediction, it can also cause noise at certain levels. In the BN, the dependence between two associated biophysical factors can be described by the strength of the edge/arc between them, which is measured from score gain/loss caused by this edge/arc's removal [24]. In general, the larger strength value represents stronger dependency, and BN structure can be adjusted from selecting different edge strength thresholds. Thus, the properties of edge strength provide an opportunity to discover the relative importance among the predictors in the BN in terms of RP2 prediction. By increasing the threshold of edge strength in a stable BN, an edge with relatively weak strength will disappear. A node without any descendant in the BN is considered as a leaf node, and it has the highest priority to be eliminated for the improvement of RP2 prediction according to the trade-off between variance reduction and loss of information. This is the basic idea of predictor pruning in our BN structure learning. Therefore, the presence of any predictor in the BN is determined by whether it is part of the extended MB, satisfies biophysical rules, and its relative strength within the network to predict RP2.

Cross validation is a technique to assess how a statistical model will generalize to an independent dataset. The Receiver Operating Characteristic (ROC) curve shows the trade-offs between true