Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk

Sangkyu Lee, Norma Ybarra, and Krishinima Jeyaseelan
Medical Physics Unit, McGill University, Montreal, Quebec H3G1A4, Canada

Sergio Faria, Neil Kopek, and Pascale Brisebois
Department of Radiation Oncology, Montreal General Hospital, Montreal, H3G1A4, Canada

Jeffrey D. Bradley and Clifford Robinson
Radiation Oncology, Washington University School of Medicine in St. Louis, St. Louis, Missouri 63110

Jan Seuntjens and Issam El Naqa
Medical Physics Unit, McGill University, Montreal, Quebec H3G1A4, Canada

(Received 9 August 2014; revised 4 February 2015; accepted for publication 26 February 2015; published 17 April 2015; corrected 23 April 2015)

Purpose: Prediction of radiation pneumonitis (RP) has been shown to be challenging due to the involvement of a variety of factors including dose–volume metrics and radiosensitivity biomarkers. Some of these factors are highly correlated and might affect prediction results when combined. Bayesian network (BN) provides a probabilistic framework to represent variable dependencies in a directed acyclic graph. The aim of this study is to integrate the BN framework and a systems’ biology approach to detect possible interactions among RP risk factors and exploit these relationships to enhance both the understanding and prediction of RP.

Methods: The authors studied 54 nonsmall-cell lung cancer patients who received curative 3D-conformal radiotherapy. Nineteen RP events were observed (common toxicity criteria for adverse events grade 2 or higher). Serum concentration of the following four candidate biomarkers were measured at baseline and midtreatment: alpha-2-macroglobulin, angiotensin converting enzyme (ACE), transforming growth factor, interleukin-6. Dose-volumetric and clinical parameters were also included as covariates. Feature selection was performed using a Markov blanket approach based on the Koller–Sahami filter. The Markov chain Monte Carlo technique estimated the posterior distribution of BN graphs built from the observed data of the selected variables and causality constraints. RP probability was estimated using a limited number of high posterior graphs (ensemble) and was averaged for the final RP estimate using Bayes’ rule. A resampling method based on bootstrapping was applied to model training and validation in order to control under- and overfit pitfalls.

Results: RP prediction power of the BN ensemble approach reached its optimum at a size of 200. The optimized performance of the BN model recorded an area under the receiver operating characteristic curve (AUC) of 0.83, which was significantly higher than multivariate logistic regression (0.77), mean heart dose (0.69), and a pre-to-midtreatment change in ACE (0.66). When RP prediction was made only with pretreatment information, the AUC ranged from 0.76 to 0.81 depending on the ensemble size. Bootstrap validation of graph features in the ensemble quantified confidence of association between variables in the graphs where ten interactions were statistically significant.

Conclusions: The presented BN methodology provides the flexibility to model hierarchical interactions between RP covariates, which is applied to probabilistic inference on RP. The authors’ preliminary results demonstrate that such framework combined with an ensemble method can possibly improve prediction of RP under real-life clinical circumstances such as missing data or treatment plan adaptation. © 2015 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution 3.0 Unported License. [http://dx.doi.org/10.1118/1.4915284]

Key words: Bayesian network, radiation pneumonitis, NTCP, biomarker, ensemble learning

1. INTRODUCTION

Radiotherapy (RT) is considered a standard of care for medically inoperable locally advanced nonsmall-cell lung cancer (NSCLC). Despite rapid advances in radiation delivery techniques during the last two decades, prognosis of NSCLC remains poor, with 5-yr survival stalled at 18.2%. RT dose escalation is excessive normal tissue toxicity and especially the risk of radiation pneumonitis (RP). In this regard, accurate prediction of RP risk may be useful for cancer cure as well as enhancing the quality of life for patients receiving radiation treatment.

Current RP prediction models in clinical use rely almost entirely on dosimetric parameters such as average dose to lung or percentage of irradiated lung volume. However, it has been reported that patients display heterogeneity in normal tissue...
radiosensitivity given identical radiotherapy regimen. Identifying over- or under-responding patients has been attempted on the basis of underlying biological reactions responsible for the endpoint of interest. Fleckenstein et al. describe the pathogenesis of radiation-induced lung disease as multiple inter-reacting cellular activities such as hypoxia, fibrogenesis, inflammation, and angiogenesis. This theory has been corroborated by several independent studies that reported association between RP risk and biomarkers for those processes [transforming growth factor (TGF)-β1, interleukin (IL)-6, α2-macroglobulin (α2M), angiotensin converting enzyme (ACE)]. However, significance of those proteins as an individual predictive biomarker is still debatable, and multivariate analysis on those biomarkers for prediction has yet to be investigated.

Some investigators combined dosimetric and clinical RP risk factors through multivariate logistic modeling to enhance the predictive performance of univariate models. However, logistic regression is limited in its flexibility to embrace dependence among multiple features. There is growing emphasis from systems’ biology perspective, as proposed by a few authors, to better understand structures of dependence between biophysical variables such as hierarchy or modulation. We propose Bayesian network (BN) as a framework to render such arrangements possible. BN is a graphical model designed for modeling joint probability distribution among random variables. Calculation of joint probability is greatly facilitated by conditional independence relationships encoded in a directed acyclic graph (DAG). In this way, we can derive a probabilistic classifier from BN by estimating the probability of a class conditional on the covariates with known values. A DAG also provides visual representation of causal relationships, which is appealing to end-users who wish to interpret inference results intuitively.

A DAG can either be specified by experts or learned from observational data. Learning DAGs from data is known to be computationally complex due to super-exponentially growing possibilities of possible graphs with the number of variables. Such complexity of the model poses a big challenge to many model selection schemes which select the single best-fitting model based on a predefined score function. However, the schemes do not always lead to an optimal classification on unseen data due to the way that the score function is designed. An alternative to model selection is retaining multiple models in an ensemble and taking the average of prediction results by the ensemble members weighted by their posterior probabilities. This approach, called Bayesian model averaging, has been shown experimentally to improve performance of complex models, including the multivariate RP prediction models by Das et al.

Bayesian network has been in use for radiotherapy outcome prediction, mainly for tumor control. To our knowledge, it has not been applied to radiation induced toxicity and especially RP where univariate predictors achieved limited success. In this paper, we intended to elaborate on the methods of BN ensemble learning as a proof of concept that could in the future improve risk prediction and also generate hypotheses on the pathophysiology of radiation-induced diseases. Background materials pertaining to the presented methodology are summarized in the supplementary material.

2. MATERIALS AND METHODS

2.A. Patient cohort

Fifty-four stage III NSCLC patients were included in this study according to the following criteria: (1) received standard 2-Gy-per-fraction 3D conformal radiotherapy with curative intent, (2) have no history of previous lung irradiation, and (3) baseline Karnofsky performance status (KPS) of equal to or greater than 70. Chemotherapy was also given neoadjuvantly or concurrently except two patients. Blood samples for biomarkers were first acquired on the CT simulation day (pretreatment) for baseline and the 15th fraction (midtreatment). Postradiotherapy toxicity was quantified using common toxicity criteria for adverse events (CTCAE) version 3. Incidence of RP, classified as CTCAE toxicity grade 2 or higher, was reported for 35% (19/54) of the patients. Detailed cohort characteristics can be seen in Table I.

2.B. Candidate variable list

We collected from blood samples and archived radiation oncology database all the information on candidate RP covariates with reported association in the literature (see Table II for references). From each blood sample, serum concentration of the following candidate RP biomarkers were measured by enzyme-linked immunosorbent assay (ELISA):

Table I. Patient cohort characteristics.

| Patient count (%) | Cohort size 54 |
|-------------------|---------------|
| RP grade          |               |
| 0                 | 34 (63)       |
| 1                 | 1 (2)         |
| 2                 | 11 (20)       |
| 3                 | 6 (11)        |
| 5                 | 2 (4)         |
| 2 and above       | 19 (35)       |
| Gender            |               |
| Male              | 23 (43)       |
| Female            | 31 (57)       |
| Prescription dose (Gy) | Median 60.0 |
| Range             | 56.0–70.0     |
| Mean lung dose<sup>a</sup> | Median 15.0 |
| Range             | 4.8–24.0      |
| Chemotherapy      |               |
| No                | 2 (4)         |
| Concurrent        | 20 (37)       |
| Neoadjuvant       | 1 (2)         |
| Neoadjuvant + concurrent | 29 (54)       |
| Adjuvant          | 2 (4)         |

<sup>a</sup>Heterogeneity-corrected, PTV excluded.
TABLE II. Candidate RP covariates. The italicized variables were included into the BN model via KS filtering.

| Variable       | Odds ratio       | Selection frequency |
|----------------|------------------|---------------------|
| a2M_pre (Ref. 11) | 0.58 (0.30, 1.10) | 0.321               |
| a2M_ratio (Ref. 11) | 1.68 (0.93, 3.01) | 0.545               |
| IL6_pre (Ref. 10) | 1.89 (0.81, 4.40) | 0.404               |
| IL6_ratio (Ref. 10) | 1.33 (0.76, 2.34) | 0.3                 |
| ACE_pre (Ref. 12) | 0.76 (0.42, 1.38) | 0.205               |
| ACE_ratio (Ref. 12) | 2.17 (1.01, 4.67) | 0.382               |
| TGFβ1_pre (Ref. 8) | 0.92 (0.52, 1.63) | 0.357               |
| TGFβ1_ratio (Ref. 8) | 0.99 (0.56, 1.74) | 0.204               |
| MLD (Ref. 42) | 1.35 (0.75, 2.24) | 0.254               |
| MHD (Ref. 44) | 2.46 (1.26, 4.83) | 0.696               |
| V20 (Ref. 43) | 1.32 (0.74, 2.36) | 0.647               |
| V30 (Ref. 43) | 1.60 (0.88, 2.93) | 0.207               |
| PTV volume | 0.99 (0.56, 1.74) | 0.21                |
| PTVCOMSI (Ref. 25) | 0.63 (0.35, 1.14) | 0.613               |
| Age (Ref. 41) | 0.95 (0.54, 1.67) | 0.366               |
| Smoking (Ref. 41) | 1.21 (0.68, 2.13) | 0.289               |

*Bootstrap frequency of 16 RP covariates selected by the KS variable filtering at a model order 6.

TGF-β1, IL-6, ACE, and a2M. In addition, the following dose-volumetric parameters were extracted from RT treatment plans based on Anisotropic Analytical Algorithm (AAA) heterogeneity correction: mean lung dose (MLD), mean heart dose (MHD), percentage volume of lung receiving 20/30 Gy or less (V20/V30), PTV volume, and superior–inferior position of PTV normalized to apical–basal extent of lung (PTVCOMSI).25 In order to account for dosimetric uncertainty due to tumor motion, PTV volume was excluded from lung dose calculation. In summary, the raw dataset contained the RP incidence and 16 RP covariates in the four main categories shown below:

- Baseline concentration (pre) of TGF-β1, IL-6, ACE, and a2M.
- Midtreatment concentration (ratio) of TGF-β1, IL-6, ACE, and a2M, taken as a percentile increase from baseline.
- RT plan parameters: MLD, MHD, V20, V30, PTV volume, PTVCOMSI.
- Clinical patient information: age, smoking status.

2.C. Data preprocessing

For Bayesian network modeling, all the covariates were discretized into two bins (high/low) because finer discretization led to sparsity in conditional probability estimation. The binning was performed using the K nearest neighbor clustering algorithm40 which grouped the data points into two clusters based on a distance (Fig. 1). The Koller–Sahami (KS) variable filter26 was used for reducing the number of covariates. The KS filter iteratively removes the variables that give little or no additional information (entropy) to a class in the presence of other variables (blankets) (see supplementary material for its theoretical basis). For our application, a blanket size of one was chosen because larger size resulted in inaccurate entropy values in our data. The filtering was implemented in the following three sequential steps: the KS filter shrunk the data down to two variables (the smallest number allowed with a blanket size = 1) and the change in the entropy of the removed along the 14 elimination rounds was recorded. The optimal dimensionality of data was defined as the number of variables with the cross-entropy higher than the value of a dummy variable filled with random values. Then, the KS filter was reapplied to the full dataset down to the determined dimensionality. This was repeated in 1000 bootstrap samples and the variables that were selected most frequently were retained for the next steps.

2.D. Bayesian network learning

The Metropolis–Hastings Markov chain Monte Carlo (MCMC) sampling for Bayesian network graphs,20,39 implemented in the Bayesian network toolbox (BNT),27 was
chosen as a graph searching algorithm. An acceptance ratio for drawing a Markov chain sample was determined by the ratio of a marginal likelihood graph score, \(P(G|D)\), and a proposal distribution \(Q(G_{\text{new}}; G_{\text{old}})\) between the new graph sample \(G_{\text{new}}\) and the last sample in the chain \(G_{\text{old}}\):

\[
A = A_1 A_2, \quad A_1 = \frac{P(G_{\text{new}}|D)}{P(G_{\text{old}}|D)}, \quad A_2 = \frac{Q(G_{\text{old}}; G_{\text{new}})}{Q(G_{\text{new}}; G_{\text{old}})} = \frac{N(G_{\text{old}})}{N(G_{\text{new}})}.
\]

A uniform distribution \(Q(G_{\text{new}}; G_{\text{old}}) = 1/N(G_{\text{old}})\) was used for the proposal density, where \(N(G_{\text{old}})\) is the number of the valid graphs that can be created from \(G_{\text{old}}\) satisfying the conditions given below:

- Should be able to be created from \(G_{\text{old}}\) by modification on a single edge (addition of a new edge or deletion/reversal of an existing edge).
- The graph after modification has to satisfy the following constraints:
  1. No loop is formed by edges (acyclicity).
  2. The number of parents for any node is no more than three (maximum fan-in).
  3. Every edge is in a causal direction (causality prior).

A causality prior imposes categorical restrictions on the presence or the direction of edges between nodes in order to reduce the search space. The allowed connections between the four categories of RP covariates (see Sec. 2) and a RP node (Fig. 2) are compatible to temporal order (e.g., baseline biomarkers \(\rightarrow\) midtreatment biomarkers \(\rightarrow\) RP) or cause–effect relationship (e.g., radiation \(\rightarrow\) midtreatment biomarkers).

Twenty-five Markov chains were created with random initial graphs and grown until convergence. Convergence of the chains was checked at every 10 000 MCMC samples. At each checkpoint, the samples were histogrammed into a graph posterior distribution, which consist of distinct graphs in the chains and the frequencies of their occurrence. Convergence was judged by sufficient consistency in the posterior distribution over time. After the convergence, the first 10 000 samples in the chains were discarded due to strong correlation with arbitrary initial condition (burn-in period) and did not contribute to estimating posterior distribution. For every graphs in the posterior, maximum \(a\) posteriori (MAP) CPTs for parent–child pairs in the DAG were learned from data with equivalent sample size \(\alpha = 1\). The expectation–maximization (EM) algorithm was applied in conjunction with MAP in order to handle missing data which constituted 9.8% of the data.

2.E. Ensemble Bayesian network classifier

A Bayesian network structure with a complete graph and CPTs is capable of making inference on the probability of RP. The inference was performed by the junction tree algorithm implemented in the BNT\(^{27}\) which is capable of dealing with variables with an unknown value in a query. Probability of RP was computed for each Bayesian network graph in an ensemble which consisted of the MCMC graph samples with \(N\) highest posterior probability \(P(i|D)\). These probability estimates were averaged over the ensemble weighted by the posterior probability of the graphs to yield the final estimate of the RP probability,

\[
P(RP) = \frac{\sum_{i=1}^{N} P_i(RP) P(i|D)}{\sum_{i=1}^{N} P(i|D)}.
\]

Predictive performance of the BN ensemble model was quantified by and receiver operating characteristics (ROCs) and a reliability plot. An ROC curve plots true positive rates against false positive rates while varying classification threshold on the estimated \(P(RP)\). The measured ROC metrics include the area under the ROC curve (AUC), sensitivity, specificity, and accuracy at the optimal operating threshold that maximizes the sum of sensitivity and specificity according to Youden’s J statistics. While ROC measures discriminative ability of a model as a classifier, a reliability plot shows accuracy of probability estimates. In order to create a reliability plot, the patients were distributed into five risk groups based on the \(P(RP)\) computed by the BN models and their actual risk (an event rate in each group) was plotted against the modeled risk (an average \(P(RP)\) in each group).

2.F. Bootstrap validation of the Bayesian network model

Graphical features and prediction power of the BN ensemble were validated in 200 nonparametric bootstrap datasets, which were intended to test robustness of the BN learning methods to limited size of data. Confidence in the presence of a certain link in a graph \((f)\), denoted as \(P(f)\), was estimated from \(M\) bootstrap replicates according to the following formula [Eq. (3)]:\(^{28}\)

\[
P(f) = \sum_{i=1}^{M} \Pr(f|D_i),
\]

where \(\Pr(f|D_i)\) denotes the probability of detecting \(f\) from Bayesian network graph training on a bootstrap replicate \(D_i\) and can be approximated by the posterior-weighted average occurrence of the link in the ensemble \(\{G\}\).\(^{29}\)
\[
\Pr(f|D) \sim \frac{\sum_{G \in \mathcal{G}} \Pr(G|D_i)f(G)}{\sum_{G \in \mathcal{G}} \Pr(G|D_i)}.
\tag{4}
\]

Classification accuracy of the BN model was also validated using the same 200 bootstrap replicates with the 0.632+ bootstrap error calculation method.\textsuperscript{30}

2.G. Comparison with other RP predictors

Predictive performance of the BN ensemble model was compared with univariate predictors in the dataset and a multivariate logistic regression predictor which was built on the covariates selected by the KS filter. For these predictors, unlike for the BN modeling, the input data retained their continuous scale after standard normalization. Missing data were filled in by \(K\) nearest neighbor imputation prior to the training. In order to minimize overfitting, a \(L^2\) regularizing term was added to a mean-square error for the objective function. The coefficient of the regularization term was tuned by 10-fold cross validation repeated 100 times with randomly assigned folds. ROC metrics using these predictors on the 200 bootstrap replicates were compared with the results from the BN model in a pairwise fashion by the Wilcoxon’s signed rank test with \(\alpha = 0.05\).

3. RESULTS

3.A. Variable selection

The KS filtering on the full dataset resulted in the optimal dimensionality of six (Fig. 3) and the following chosen variables: midtreatment \(\alpha^2M\), (2) pretreatment IL-6, (3) midtreatment ACE, (4) MHD, (5) \(V_{20}\), and (6) PTVCOMSI (Table II). The selected variables displayed a varying degree of univariate correlation with RP, with \(V_{20}\) at the weakest (odds ratio = 1.32) and MHD at the strongest (odds ratio = 2.46).

3.B. Bayesian network learning with MCMC

Figure 4 demonstrates that likelihood of the sampled graphs rose from the initial suboptimal locations to reach an oscillatory equilibrium shortly after 1000 runs. However, equilibrium on estimating posterior distribution of graphs did not occur until at least 50 000 runs (Fig. 5). The convergence of the posterior distribution was attributed to a restricted search space imposed by the causality prior, as the chain failed to converge within reasonable time in the absence of such conditions.

3.C. Confidence in BN graphical features

When the ensemble size of 200 was used, a median confidence level of the 23 possible connections allowed by the causal prior was 0.27. Ten out of 23 achieved the confidence level higher than chance (0.29)\textsuperscript{31} (Table III, Fig. 6). The highest value was recorded for the connection from ACE\_ratio to RP (0.93).

3.D. Prediction of RP risk

Predictive power of the Bayesian network method was compared with that of individual biomarkers and dosimetric RP models in the literature.\textsuperscript{32–34} As seen by changes in AUC and specificity (Fig. 8), prediction with the BN method greatly improved as the ensemble size grew up to 200 beyond which the increase tailed off. For any size of the ensemble, the AUC for the BN was larger than any univariate classifiers (Fig. 7). The highest AUC from univariate biomarkers and dosimetric models was recorded by midtreatment ACE (0.66) and MHD (0.69), which was significantly lower than the BN model at an ensemble size \(N = 200\) (0.83). Given the same set of variables, multivariate logistic regression was shown to be less effective in classification than BN at any ensemble size, as seen from significantly lower AUC and specificity at the optimal operating threshold (Fig. 7, Table IV).

![Fig. 3. Cross entropy of the variables removed at each round of KS backward elimination. A red line indicates the range of cross entropy from a dummy variable. Error bars: 95% confidence interval from bootstrapping.](image1)

![Fig. 4. Change in likelihood graph score and acceptance rate over 60000 MCMC iterations shown on log scale. Twenty-five chains with random initialization were averaged.](image2)
Accuracy of probability estimates of the multivariate models was evaluated via a reliability plot (Fig. 9). The accuracy was measured as a goodness of fit to a perfect probability estimate which corresponds to a diagonal line in the plot. Similarly to ROC metrics, the Bayesian network model with larger ensemble size was shown to give more accurate estimates of RP risks.

In order to examine the ability of the Bayesian network ensemble to perform probabilistic inference under missing data, the probability of RP was estimated in the absence of intratreatment biomarker data. AUC values decreased significantly as a result, but with varying degrees depending on ensemble sizes. For larger ensemble sizes above 200, the reduction in AUC was smaller (0.03) than for sizes less than 50 (0.05) (Fig. 8).

4. DISCUSSION

High input dimensionality is certainly one of the biggest challenges in multivariate disease modeling, especially in radiotherapy induced sequelae where dosimetric, biological, and patient-specific clinical parameters all contribute to extra complexity. Dependence between risk factors is inevitable in radiotherapy informatics and has been neglected by many multivariate methods such as logistic regression in the name of the “black box” approach. Naive independence assumption of such models may lead to increased bias in probability estimation and consequently poor inference with unseen data. Alternatively, this variety of factors could be structured into systems of different hierarchical levels between which connections could be established by fusing observation with prior knowledge. Bayesian network provides the statistical platform in which this idea of systems’ biology can be implemented and applied for disease prediction. However, the amount of current expert knowledge in this area does not fully cover all the putative RP risk factors. Thus, feasibility of learning BN structures from radiation oncology data needs to be addressed.

Table III. Bootstrap estimated confidence of the connections shown in Fig. 6.

| Parent      | Child      | Confidence |
|-------------|------------|------------|
| ACE_ratio   | RP         | 0.93       |
| MHD         | RP         | 0.90       |
| α2M_ratio   | RP         | 0.70       |
| IL6_pre     | V20        | 0.58       |
| PTVCOMSI    | MHD        | 0.53       |
| IL6_pre     | ACE_ratio  | 0.43       |
| MHD         | ACE_ratio  | 0.40       |
| MHD         | α2M_ratio  | 0.37       |
| ACE_ratio   | α2M_ratio  | 0.36       |
| IL6_pre     | PTVCOMSI   | 0.32       |

Fig. 6. Variables connected by directed edges with a confidence level higher than random. Edge thickness is proportional to its confidence level which is also written on the edge (Table III). Arrow-headed and bar-headed edges are assigned to positive and negative correlation, respectively.
Bayesian network is intrinsically a complex modeling technique with a multitude of parameters. Several measures were taken in training and testing the BN prediction model to address the issue of overfitting: a causality prior and a fan-in bound imposed restrictions on the number of possible graphs to narrow down the search space. In addition, marginal likelihood as a graph scoring function penalizes larger number of edges and helps to find a balance between the complexity and fitness of data. Generalizability of the BN classifier was validated in a 0.632+ bootstrap setting which was shown to be the closest to the true population estimate in small training sample size and high model orders. However, even these efforts do not remove the uncertainty of a single BN model learned from a modest size dataset. We observed from MCMC simulation the presence of multiple high-scoring graphs also known as likelihood equivalence. We addressed this issue by adopting a Bayesian approach which embraces an ensemble of models rather than a single one. Having a bag of models compromises graphical interpretation of a single model but overcomes the uncertainty in model selection and improves the performance on unseen data, which was shown in our cohort by larger AUC with increasing ensemble size.

We observed superiority of the Bayesian network method over logistic regression in classifying RP events (Fig. 7). The difference could be attributed to a probabilistic approach of Bayesian network. It models a joint probability distribution \( p(x, y) \) for covariates \( x \) and a class \( y \) and for classification computes \( p(y|x) \) using the Bayes’ rule, in contrary to logistic regression which is limited to computing \( p(y|x) \). The advantage of the former, already reported by Ng and Jordan in several datasets with a modest size, might explain more accurate RP probability estimates by BN models in our data (Fig. 9). Also, the probabilistic nature of BN enables inference under incomplete data. For nonprobabilistic models such as logistic regression, missing data have to be filled in by imputation, which inevitably corrupts the integrity of data. Furthermore, we were looking into capitalizing on this feature to attempt “hastened” prediction before receiving midtreatment biomarkers. Although some reduction in AUC was observed, the model to a certain degree compensated for the missing information by using graphical relationships and parameters obtained with complete training data. This is due to the role that ensemble plays (Fig. 8) and its resulting high AUC values (0.76–0.81) compared to MHD (0.70), which was the only variable connected to RP other than midtreatment biomarkers (Fig. 6).

A Markov chain Monte Carlo method was used for exploring a model space and approximating posterior probability. As seen from Fig. 5, the graph samples from MCMC showed power law-like distribution which consists of a small number of high-posterior graphs and a long tail of low posterior. It turned out that only the graphs around 200, constituting the peak and a small part of the “tail” at cumulative probability less than 30%, are necessary for the optimal RP classification. This result agrees well with previous studies showing that high posterior models are indeed beneficial for classification tasks as well. We also derived from those graphs the confidence levels of graph features through repeated training in bootstrap replicates. Edges with high confidence might imply existence of novel causal relationships. Although we did not conduct any independent bench experiment to validate the links, this feature of the BN model is definitely an advantage in generating new hypotheses compared to many black-box type multivariate models.

Precautions need to be taken, however, in interpreting the results of this study for clinical application of the presented
Bayesian network methodology. Small number of RP events due to modest dataset size and low-risk nature implies higher uncertainty in predicting positive instances than the negatives. This was demonstrated in Fig. 9 showing worse precision of probability estimates for higher risk patients. Another limitation, inherent to BN structure learning, is that the connections between variables that we identified from BN graphs may not always indicate direct causality. Although noncausal links were ruled out during the structure learning, it is possible that a pair of variables without any direct causality is seemingly correlated through any confounding factor hidden from the model. Due to computational burden from MCMC and a hazard of overfitting under a limited sample size, input dimension had to be tightly constrained by use of variable selection. The KS filtering scheme was shown to be effective in removing the variables that do not give. In this study, it was useful in reducing the dimensionality of a dose-volumetric domain that contained highly correlated variables. However, this type of filtering might have dropped the variables that could have been more useful in encoding causal relationships.

Future development of this study will address the effect of different fractionation schemes on variable selection, types of biophysical interactions identified by BN, and RP prediction by applying the described method on a SBRT cohort. We are also developing a method for systematically encoding domain knowledge in radiobiology into the graph learning. In addition to the causality constraints, this is expected to identify true biological relationships with high confidence and thus improve the generalizability of the model. Finally, although the model was validated internally using 0.632+ bootstrap, an external source of data with matching characteristics would be necessary to add more clinical relevance to our findings.

![Fig. 9. Reliability plot of three probabilistic classifiers. Numbers in parentheses indicate the r² value with respect to an ideal probability estimator (dashed red line). Standard errors, shown in error bars, were computed from binomial distribution (actual risk) or quadrature summation of variance of probability scores in a group and bootstrap-estimated model uncertainty (predicted risk).](image-url)
5. CONCLUSION

We applied a Bayesian network framework in conjunction with Bayesian statistics for constructing interaction graphs of biological, dosimetric, and clinical covariates for radiation pneumonitis. Markov chain Monte Carlo sampled high posterior graphs which were used as an ensemble to estimate radiation pneumonitis risk. We have shown that a certain size of the ensemble was sufficient to perform optimal classification. Bootstrap-validated predictive power of the Bayesian network ensemble was superior to any univariate predictors or multivariate logistic regression. Statistical confidence in graph features in the ensemble obtained by bootstrapping can potentially identify novel biophysical relationships. After validation on larger dataset, the presented modeling strategy could be useful for estimating normal tissue complication probability (NTCP) for various endpoints.

ACKNOWLEDGMENTS

The authors acknowledge the consults from Dr. Jung Hun Oh and Dr. Natalie Japkowicz. The authors also thank Dr. Christina Speirs for providing them a part of clinical data. The computational work was enabled in part by computer resources provided by WestGrid (www.westgrid.ca) and Calcul Quebec (http://www.calculquebec.ca/). This research was funded by the Canadian Institute of Health Research (CIHR) grant MOP-114910, Excellence in Radiation Research for the 21st Century (CIHR, Canada), and the CREATE Medical Physics Research Training Network grant (NSERC, Canada).

---

2429 Lee et al.: Multivariate BN ensemble to predict RP risk
B. Efron and R. Tibshirani, “Improvements on cross-validation: The .632+ bootstrap method,” J. Am. Stat. Assoc. 92(438), 548–560 (1997).

M. Scutari, “On the prior and posterior distributions used in graphical modeling,” Bayesian Analysis 8(3), 505–532 (2013).

G. J. Kutcher and C. Burman, “Calculation of complication probability factors for non-uniform normal tissue irradiation: The effective volume method gerald,” Int. J. Radiat. Oncol., Biol., Phys. 16(6), 1623–1630 (1989).

A. Niemierko, “Reporting and analyzing dose distributions: A concept of equivalent uniform dose,” Med. Phys. 24(1), 103–110 (1997).

J. D. Bradley, A. Hope, I. El Naqa, A. Apte, P. E. Lindsay, W. Bosch, J. Matthews, W. Sause, M. V. Graham, and J. O. Deasy, “A nomogram to predict radiation pneumonitis, derived from a combined analysis of rtog 9311 and institutional data,” Int. J. Radiat. Oncol., Biol., Phys. 69(4), 985–992 (2007).

N. Friedman and D. Koller, “Being bayesian about network structure—Bayesian approach to structure discovery in bayesian networks,” Mach. Learn. 50, 95–125 (2003).

D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning (The MIT press, MA, 1999).

B. Sahiner, H. Chan, and L. Hadjiiiski, “Classifier performance prediction for computer-aided diagnosis using a limited dataset,” Med. Phys. 35(4), 1559–1570 (2008).

A. Ng and M. Jordan, “On Discriminative vs Generative classifiers: A comparison of logistic regression and naive Bayes,” in Advances in Neural Information Processing Systems (NIPS) (The MIT Press, 2001), Vol. 14.

G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Application (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007).

B. Parashar, A. Edwards, R. Mehta, M. Pasmanter, A. G. Wernicke, A. S Abbas, R. S. Kerestez, D. Nori, and K. S. Chao, “Chemotherapy significantly increases the risk of radiation pneumonitis in radiation therapy of advanced lung cancer,” Am. J. Clin. Oncol. 34, 160–164 (2011).

S. L. Kwa, J. V. Lebesque, J. C. Theuws, L. B. Marks, M. T. Munley, G. B lentel, D. Oetzel, U. Spahn, M. V. Graham, R. E. Drzymala, J. A. Purdy, A. S. Lichter, M. K. Martel, and R. K. Ten Haken, “Radiation pneumonitis as a function of mean lung dose: An analysis of pooled data of 540 patients,” Int. J. Radiat. Oncol., Biol., Phys. 42(1), 1–9 (1998).

F. M. Kong, J. A. Hayman, K. A. Griffith, G. P. Kalemkerian, D. A nenberg, S. Lyons, A. Turrisi, A. Lichter, B. Fraass, A. Eisbruch, T. S. Lawrence, and R. K. Ten Haken, “Final toxicity results of a radiation-dose escalation study in patients with non-small-cell lung cancer (NSCLC): Predictors for radiation pneumonitis and fibrosis,” Int. J. Radiat. Oncol., Biol., Phys. 65(4), 1075–1086 (2006).

E. X. Huang, A. J. Hope, P. E. Lindsay, M. Trovo, I. El Naqa, J. O. Deasy, and J. D. Bradley, “Heart irradiation as a risk factor for radiation pneumonitis,” Acta Oncol. 50(1), 51–60 (2011).

See supplementary material at http://dx.doi.org/10.1118/1.4915284 for materials on the machine learning concepts pertaining to the presented methodology.