International Journal of
Radiation Oncology
biology • physics

www.redjournal.org

Physics Contribution

Integrating Multiomics Information in Deep Learning Architectures for Joint Actuarial Outcome Prediction in Non-Small Cell Lung Cancer Patients After Radiation Therapy

Sunan Cui, PhD,[*,†] Randall K. Ten Haken, PhD,[*] and Issam El Naqa, PhD[*]

[*]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan; and [†]Applied Physics Program, University of Michigan, Ann Arbor, Michigan

Received Jul 17, 2020, and in revised form Nov 10, 2020. Accepted for publication Jan 23, 2021.

Purpose: Novel actuarial deep learning neural network (ADNN) architectures are proposed for joint prediction of radiation therapy outcomes—radiation pneumonitis (RP) and local control (LC)—in stage III non-small cell lung cancer (NSCLC) patients. Unlike normal tissue complication probability/tumor control probability models that use dosimetric information solely, our proposed models consider complex interactions among multiomics information including positron emission tomography (PET) radiomics, cytokines, and miRNAs. Additional time-to-event information is also used in the actuarial prediction.

Methods and Materials: Three architectures were investigated: *ADNN-DVH* considered dosimetric information only; *ADNN-com* integrated multiomics information; and *ADNN-com-joint* combined RP2 (RP grade $\geq$2) and LC prediction. In these architectures, differential dose-volume histograms (DVHs) were fed into 1D convolutional neural networks (CNN) for extracting reduced representations. Variational encoders were used to learn representations of imaging and biological data. Reduced representations were fed into Surv-Nets to predict time-to-event probabilities for RP2 and LC independently and jointly by incorporating time information into designated loss functions.

Results: Models were evaluated on 117 retrospective patients and were independently tested on 25 newly accrued patients prospectively. A multi-institutional RTOG0617 data set of 327 patients was used for external validation. *ADNN-DVH* yielded cross-validated c-indexes (95% confidence intervals) of 0.660 (0.630-0.690) for RP2 prediction and 0.727 (0.700-0.753) for LC prediction, outperforming a generalized Lyman model for RP2 (0.613 [0.583-0.643]) and a generalized log-logistic model for LC (0.569 [0.545-0.594]). The independent internal test and external validation yielded similar results. *ADNN-com* achieved an even better performance than *ADNN-DVH* on both cross-validation and independent internal test. Furthermore,

Corresponding author: Sunan Cui, PhD; E-mail: sunan@umich.edu

This work was partly supported by grants from National Institutes of Health (NIH) grants P01 CA059827 and R01-CA233487.

Disclosures: The authors have no conflicts of interest to disclose.

Data Sharing Statement: Part of these data are publicly available (RTOG0617). Institutional research data are not available at this time.

Supplementary material for this article can be found at https://doi.org/10.1016/j.ijrobp.2021.01.042.

Acknowledgments—The authors acknowledge that the RTOG0617 data set was made available by NCT00533949-D1 from the NCTN Data Archive of the National Cancer Institute's (NCI) National Clinical Trials Network (NCTN). Data were originally collected from clinical trial NCT number NCT00533949 (A randomized phase III comparison of standard-dose [60 Gy] versus high-dose [74 Gy] conformal radiotherapy with concurrent and consolidation carboplatin/paclitaxel +/- cetuximab (Ind #103444) in patients with stage IIIA/IIIB non-small cell lung cancer). All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI.

Int J Radiation Oncol Biol Phys, Vol. 110, No. 3, pp. 893−904, 2021

https://doi.org/10.1016/j.ijrobp.2021.01.042

ADNN-com-joint, which yielded performance similar to *ADNN-com,* realized joint prediction with c-indexes of 0.705 (0.676-0.734) for RP2 and 0.740 (0.714-0.765) for LC and achieved an area under a free-response receiving operator characteristic curve (AU-FROC) of 0.729 (0.697-0.773) for the joint prediction of RP2 and LC.

Conclusion: Novel deep learning architectures that integrate multiomics information outperformed traditional normal tissue complication probability/tumor control probability models in actuarial prediction of RP2 and LC.

Introduction

Outcome modeling,[1] which aims to predict individuals' radiation therapy responses (eg, toxicity and tumor control) can potentially guide personalized prescription and support decisions for patient-specific plans before or during an adaptive treatment.[2] In this study, clinical endpoints of local control (LC) and radiation pneumonitis (RP) were considered. LC is directly correlated with long-term survival[3] in lung cancer. To achieve better LC, dose escalation can be applied while being delivered safely.[4,5] However, dose escalation is restricted by RP, which can induce life-threatening side effects. Hence, modeling LC/RP will be a critical step to optimize and personalize radiation therapy treatment. Specifically, RP2 (RP $\geq$2) is considered in the study because it is a major limiting factor for dose escalation in lung cancer patients.

Currently, prevalent outcome models are commonly based on dosimetric information[6] including the Lyman model[7] for normal tissue complication probability (NTCP)[8] and log-logistic models for tumor control probability (TCP).[9,10] Specifically, in these models, a function with several free parameters is chosen to model the dependence of NTCP/TCP on radiation dose. Such models are usually based on a fixed sigmoid-shaped function (eg, cumulative Gaussian distribution (probit), logistic function). Dosimetric information other than simple metrics such as generalized equivalent uniform dose (gEUD) or dose thresholds (V20) is not used. Hence, the models may hardly reflect the complex interplay of outcome, dose distribution, and other confounding variables. In a previous study, principal component analysis (PCA) was applied to extract features from whole dose-volume histograms (DVHs) for the prediction of radiation toxicity[11] in liver and parotid gland. PCA is an unsupervised technique aiming to re-express the data such that signal is retained and noise is filtered out. Principal components (ie, extracted features in PCA) are a linear combination of original inputs, which may not be enough for revealing nonlinear associations in the prediction task. Alternatively, our neural network architectures *ADNN-DVHs* are proposed to directly learn nonlinear-morphologic characteristics from DVHs for the prediction of RP2 and LC.

It is also recognized that radiation therapy outcomes are associated with complex interactions among patient-specific information. Hence, outcome models considering physical, imaging, and biological data have continued to be of interest. Machine learning methods such as support vector machines,[12] neural networks (NNs), random forests, and Bayesian networks[13] have shown promising results in modeling multiomics outcomes. Several reviews[14-19] have discussed the limitations of conventional machine learning—based outcome models and the potential of deep learning to overcome current barriers. In conventional machine learning models, feature selection was a necessary component to reduce the dimensionality of the high-dimensional patient-specific data into a smaller candidate set for processing.[20] In our previous studies, deep learning techniques were applied to outcome modeling. Specifically, a composite architecture consisting of variational autoencoders (VAEs) and a multi-layer perceptron (MLP) for RP2,[20] and a composite architecture consisting of a 1D CNN and a MLP[21] were investigated for LC. However, these architectures, requiring simple prescreening of features, did not fully realize deep learning's potential (ie, making the prediction directly from the high-dimensional raw data set). In this study, variational encoders (VAEs)[22] and 1D convolutional neural networks (CNNs) were adopted for learning representations from biological, imaging, and dosimetric data, and reduced representations were subsequently fed into survival neural networks[23] (Surv-Net). The proposed architecture *ADNN-com,* which conducts simultaneous dimensionality reduction and prediction, can realize direct outcome prediction from a heterogeneous and high-dimensional data set.

Unlike most outcome models, which usually focus on predicting a single outcome, multi-endpoint predictions[11] can be realized in our study; that is, prediction of RP2 and LC can be simultaneously generated from a single architecture *ADNN-com-joint.* Hence, trade-offs between competing outcomes, such as RP2 and LC, can be possibly handled in our models while accounting for cross-talks which is an important step toward establishing robust outcome models and more reliable decision-support tools.

Additionally, temporal information was considered in all of our proposed outcome models. This allows the prediction of time to event in addition to classifying events, as commonly practiced in the outcome modeling literature. Conventional dose-response models, (eg, Lyman models, log-logistic models) and machine learning models (eg, support vector machines and NNs) are usually designed for binary/multiclass classifications. However, compared with a binary endpoint, which is attached to a specific follow-up time, time to toxicity/progression would leverage additional temporal information into outcome models and help provide better time-dependent decision support. Moreover, incorporation of time-to-censor information will help use the censored information, which would be discarded otherwise. In the prediction of radiation therapy response, censored data are very common because follow-ups may be missed or patients may die before the event occurs.

Some efforts have been made to incorporate time-to-event information into radiation therapy outcome models. Tucker et al[24] investigated the feasibility of a generalized Lyman model for RP risk. A log-normal distribution was fitted empirically to time-to-event data, and its cumulative distribution function was then calculated and embedded into the Lyman model as a modified factor. NNs have also been investigated for incorporating follow-up information. A Cox-nnet,[25] which introduced a Cox loss function into NNs, was applied to predict patient survival with high-throughput transcriptomic data in 10 TCGA RNA-Seq data sets. Cox proportional hazards models,[26] which assume the hazard ratio between 2 different observations is constant over time, are classical models for survival analysis. However, the proportional hazard assumption may not hold in many clinical situations. A Nnet-survival model,[23] which considers discrete-time endpoints, allows the baseline hazard probability and hazard probability's dependence on input data to vary with time. In this study, a discrete-time survival network (Surv-Net), a variant of the Nnet-survival model, was combined with dimensionality reduction architectures for actuarial prediction of RP2 and LC.

Methods and Materials

Data sets

A diagram describing training, independent internal testing, and external validation processes in this study following TRIPOD recommendations[27] is shown in Figure 1. A total of 117 institutional patients and 80 TCGA patients were included in the models' development, and 25 institutional patients and 327 RTOG0617[28] patients were used for independent internal test and external validation.

Training set

In this study, 117 patients with stage III NSCLC treated with radiation therapy under institutional review board—approved protocols at our institution were used for the development of models. Following TRIPOD criteria level 2 type a (random split), stratified 5-fold cross-validation was conducted 20 times. Among 117 patients, about three-quarters were treated with standard doses (<74 Gy at 2 Gy per fraction), and the rest were treated with a dose escalation protocol intensifying dose to 2.1 to 2.85 Gy per fraction, with a total dose up to 86 Gy. To account for the effects of different fractionation schemes, all physical dose values were converted to their 2 Gy equivalents (EQD2) using the linear-quadratic model (a/b = 4 Gy for lung voxels, a/b = 10 Gy for tumor voxels).

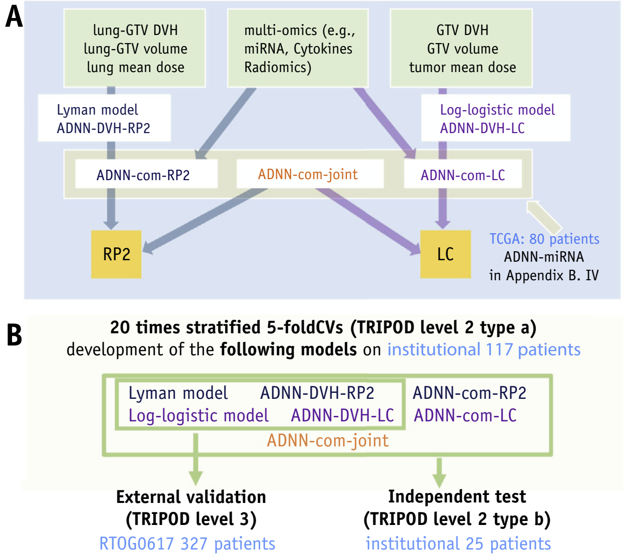


Fig. 1. Model development, independent internal test, and external validation of proposed models and normal tissue complication probability (NTCP)/tumor control probability (TCP) models. (A) The inputs and outputs of each model. (B) The scheme of model evaluations, which follows TRIPOD level 2 type a, b and TRIPOD level 3. *Abbreviations*: DVH = dose-volume histogram; GTV = gross tumor volume.

In addition to dosimetric data, biological data including levels of 30 cytokines and 60 miRNAs were collected. In addition, 43 radiomics features[29] following IBSI recommendations[30] were obtained from gross tumor volumes (GTVs) of pre- and midtreatment positron emission tomography (PET) imaging, respectively. GTVs were processed with standard uptake value and a Lloyd-Max quantization algorithm (32 gray-level)[29] before the extraction of the radiomics features. More details about the process and implementation of radiomics extraction are described in Appendix E1. A comprehensive list of the patient-specific information being considered for prediction of RP2 and LC is also presented in Appendix E1. Tumor LC was determined by physicians based on clinical, radiographic, or biopsy evidence. RP grade was classified according to the 5-grade Common Terminology Criteria for Adverse Events (CTCAE) 3.0 criterion, and RP2 was defined as RP grade ≥2.

External data sets, the cancer genome atlas (TCGA),[31] lung adenocarcinoma (LUAD),[32] and lung squamous cell carcinoma (LUSC)[33] data collections were used for transfer learning.[34] Only the data from patients who were treated with adjuvant radiation therapy, with primary tumors as treatment sites, who had more than 90 days' follow-up were kept. A total of 80 patients in TCGA with biological data were eventually included in the transfer learning task.

Testing and external validation data sets

Following TRIPOD criteria level 2 type b[27] (nonrandom split based on time), an independent prospective data set containing 25 patients with newly treated stage III NSCLC from our institution was collected for independent internal test. Following TRIPOD criteria level 3,[27] planning CT images, structure sets, plan dose, and clinical information of patients treated under the RTOG 0167 protocol[35] at a variety of institutions were analyzed and used to validate the models. Among the patients in this data set, 333 were treated with per-protocol radiation therapy treatment plans. However, 5 of 333 patients did not have available dosimetric information, and 1 patient had erroneous dosimetric information; information from 327 patients was eventually used for external validation.

Patient characteristics and endpoint information in all the data sets are summarized in Appendix E1.

Methodology

Generalized versions of Lyman[7] and log-logistic models that can account for time-to-event/censored time were applied for comparison purposes. The implementation of Lyman/log-logistic models is described in Appendix E2.

Incorporation of time-to-event/censored time in *ADNN* model

Time-to-event/censored time is incorporated into the prediction task by the deployment of Surv-Net[23] in *ADNN* models. Surv-Net, which is composed of several fully connected layers, takes the reduced representations of multiomics features as input and predicts conditional event-free probabilities through discretized time intervals. The range of discretized time intervals was determined so as to ensure the same number of events was used in each interval in the training data.

In the situation of K time intervals, the outputs can be denoted as $(P_{T_1}, P_{T_2}, \ldots, P_{T_K})$, and the log-likelihood function for an individual with failure in interval j is defined in Eq. 1.

$$l = \left(1 - P_{T_j}\right) \prod_{i=1}^{j-1} P_{T_i} \qquad \text{Eq.1}$$

The log-likelihood function for an individual without experiencing events through interval j (either censored or event-free during follow-up) is defined in Eq. 2.

$$l = \prod_{i=1}^{j} P_{T_i} \qquad \text{Eq.2}$$

The total loss function is defined as a negative average of log-likelihood function over all patients. More details about the implementation of the loss function are presented in Appendix E2.

Model *ADNN-DVH*

In *ADNN-DVH* as shown in Figure 2A, 3 blocks of 1D convolutional layers and average pooling layers were applied to differential DVHs with a bin size of 0.1 Gy. Reduced representations of DVHs were then concatenated with structure volume and mean dose, serving as inputs to Surv-Net, which consisted of 2 fully connected layers. The loss function of *ADNN-DVH* is defined in Eq. 1. (for event case) and Eq. 2.(for censored case).

Model *ADNN-com*

In *ADNN-com,* biological and imaging information is considered together with dosimetric information for the prediction of RP2 and LC. Three VAEs were applied to conduct dimensionality reduction for PET radiomics, cytokines, and miRNA data, respectively. Specifically, a trained VAE based on TCGA-LUAD and TCGA-LUSC data was used for miRNA data, which allows for more reliable and generalizable representation by transfer learning.[34] Technical details are provided in Appendix E2.

In *ADNN-com* for RP2 prediction, inputs of Surv-Net were composed of lung mean dose, lung volume, reduced representation of lung DVH, miRNA, and cytokines. For LC prediction, inputs of Surv-Net are composed of tumor mean dose, tumor volume, reduced representation of tumor DVH, PET tumor radiomics, miRNA, and cytokines. The total loss function of the *ADNN-com* architecture is composed of the sum of VAE losses and Surv-Net loss.

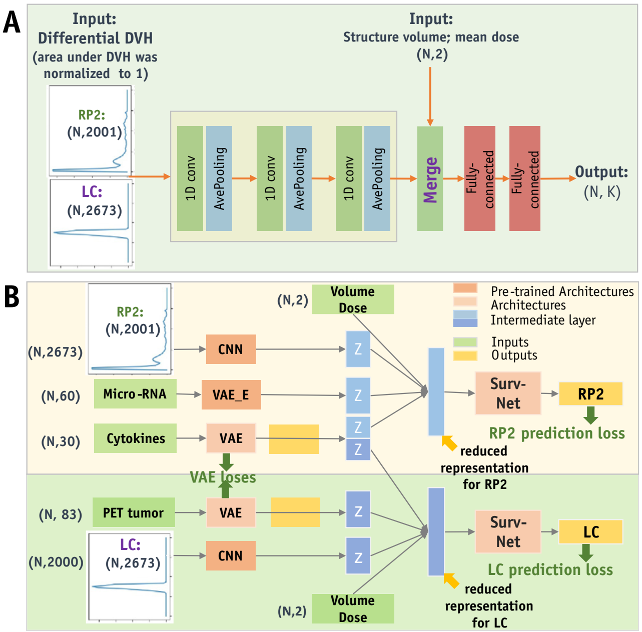


Fig. 2. Diagrams of ADNN architectures. (A) An architecture ADNN-DVH that predicts radiation pneumonitis (RP) and local control (LC) based on dosimetric information. (B) Architecture ADNN-com, which realizes the actuarial (joint) prediction of RP2 and LC. *Abbreviations*: CNN = convolutional neural network; DVH = dose-volume histogram; PET = positron emission tomography; VAE = variational encoder.

Model *ADNN-com-joint*

An *ADNN-com-joint* model as presented in Figure 2B as the realized joint prediction of RP2 and LC. It can be regarded as a combination of *ADNN-com* architectures for RP2 and LC. However, in this case, the VAEs that are applied for dimensionality reduction of cytokines and miRNA are shared between RP2 and LC. Additionally, Surv-Nets for prediction of RP2 and LC are trained simultaneously.

Grad-CAM

To gain further insights into our proposed deep learning models, gradient-weighted class activation mapping (Grad-CAM),[36] a class-discriminative localization technique, was applied to generate visual interpretation for convolutional layers as well as fully connected layers of VAEs in *ADNN-DVH* and *ADNN-com-joint.*

Grad-CAM can highlight (assign higher values to) regions in an activation map that are important for the endpoints of interest (ie, LC and RP2). A Grad-CAM entry was calculated for each convolutional layer to understand the importance of each neuron for a decision of interest. In Eq. 3, c denotes an arbitrary output (LC or RP2); $A^k \in \mathbb{R}^{u\times v}$ is the k^{th} feature map with height u and width v; α_k^c is the weight of the k^{th} feature map in discriminating output c. The weight a as shown in Eq. 4 is defined as gradients of a score for class y^c with respect to feature maps A^k of a convolutional layer followed by a global average pooling (Z is a total number of elements in the feature map). The weight a captures the importance of feature map k for a target class c (ie, LC and RP2).

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha_k^c A^k\right) \qquad \text{Eq.3}$$

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}} \qquad \text{Eq.4}$$

To visualize the fully connected layers in VAE, a modified version of Grad-CAM was adopted, as shown in Eq. 5. This version of Grad-CAM was also based on the multiplication of activation maps and the gradient of output with regard to activation maps. However, because "channels" do not exist in fully connected layers, there is no "weighted sum over channels" in this formula.

$$\left(L^c_{Grad-CAM}\right)_i = ReLU\left(\frac{\partial y^c}{\partial A_i}A_i\right) \qquad \text{Eq.5}$$

Other variations of the *ADNN* models

Some variations of *ADNN* models are also evaluated and compared with original models in the Discussion section. Specifically, a modified *ADNN-DVH* with solely DVHs (normalized by volume) as input (without dose and volume input), which helps explain the rationale of selected inputs of our *ADNN-DVH* model, is presented in Appendix E4. A modified ADNN-com model with different categories of multiomics as input, which provides insight into the relative importance of different categories of information, is described in Appendix E4. A model *ADNN-3Ddose,* which can potentially take into account spatial information, is applied to predict RP2 (Appendix E4). However, it did not achieve result comparable to *ADNN-DVH,* which is mentioned in the Discussion section.

Model implementation

Free parameters in Lyman and log-logistic models were optimized through maximum likelihood estimation by R package bbmle.[37] Architectures *ADNN-DVH, ADNN-com,* and *ADNN-com-joint* were implemented with the Python deep learning library Pytorch[38] and were trained with a NVidia Tesla V100 GPU in a high-performance computer cluster. The total training (5-fold cross-validation conducted 20 times) time of each model is roughly 2 hours. In ADNN architecture, the loss functions were optimized by ADAM methods[39] with a learning rate of 0.001. Details on the parameters used in the ADNN architectures are presented in Appendix E2.

Evaluation of performance

Harrell's c-index[40] can evaluate goodness of fit for models that produce time-dependent risk scores. Supposing a pair of patients (i, j) have risk scores $(S_i,\ S_j)$ and time to event $(T_i,\ T_j)$, with $S_i > S_j$. If $T_i < T_j$, the pair is regarded as a concordant pair, and if $T_i > T_j$, it is a discordant pair. The c-index is then defined as a ratio of concordant pairs to a sum of concordant pairs and discordant pairs. This concept can be easily adopted for Cox models, which produce a single risk score for each patient. However, because our models allow hazard probability's dependence on input data to vary with time, there is no single score (different scores in different intervals) for an individual.

Alternatively, the c-index for binary data (Eq. 8), which is based on risk score S and event L attaching to certain follow-up time τ, is adopted, and patients with censored time before τ (censored data: $d(\tau) = 0$) were excluded from the calculation. Specifically, the c-index in Eq. 8 can be regarded as the area under the receiver operating character (ROC) curve (AUC),[41] except it additionally considers time to event and censored time. In our study, because events—local progression (LP = 1; ie, LC = 0) and RP2—are relatively sparse, performance was designated to be evaluated at τ = maximal event time in each data set to cover all events; accordingly, any patients with follow-up less than τ (censored) are excluded from calculation:

$$c = \frac{\sum_{i\neq j} 1\{S_i(\tau) > S_j(\tau)\}1\{L_i(\tau)=1,\ L_j=0\}d_j(\tau)}{\sum_{i\neq j} 1\{L_i(\tau)=1,\ L_j=0\}d_j(\tau)} \qquad \text{Eq.8}$$

To further evaluate the performance of *ADNN-com,* the area under a free-response ROC (AU-FROC) curve,[42] widely used in the diagnostic classification in which cases

Table 1 Cross-validated C-index with 95% CIs and C-index in independent internal test/external validation

		RP2	LC	RP2 and LC (FROC)
Model evaluation on institutional 117 patients				
C-index (95% CI) and *P* values for comparison to analytical model				
Lyman/log-logistic		0.613 (0.583-0.643)	0.569 (0.545-0.594)	N/A
ADNN-DVH	C-index	0.660 (0.630-0.690)	0.727 (0.700-0.753)	N/A
	P value	.0293	<2.2E-16	
ADNN-com	C-index	0.691(0.661-0.722)	0.735(0.710-0.761)	0.721 (0.677-0.768)
	P value	1.305E-5	<2.2E-16	
ADNN-com-joint	C-index	**0.705 (0.676-0.734)***	**0.740 (0.715-0.765)***	**0.729 (0.697-0.773)***
	P value	**1.472E-5***	**< 2.2E-16***	
Independent internal test on 25 newly treated patients				
C-index (*P* values for comparison to analytical model)				
Lyman/log-logistic		0.588	0.573	N/A
ADNN-DVH		0.667 (0.062)	0.706 (0.045)	N/A
ADNN-com		0.683 (0.058)	0.713 (0.038)	0.687
ADNN-com-joint		**0.691 (0.051)***	**0.721 (0.029)***	**0.709***
RTOG 0617				
C-index (*P* values for comparison to analytical model)				
C-index		RP3	LC	N/A
Lyman/log-logistic		0.736	0.554	N/A
ADNN-DVH		0.762 (0.037)	0.618 (0.014)	N/A

Abbreviations: DVH = dose-volume histogram; FROC = area under a free-response receiving operator characteristic curve; LC = local control; RP = radiation pneumonitis; RTOG = Radiation Therapy Oncology Group.

* The model with best performance and its comparison with other models.

might contain 2 or more task-related lesions, was adopted. Similar to ordinary ROC curves, the ordinate in a FROC plot is true positive rate; however, this true positive rate is defined over all endpoints. The abscissa of the FROC plot is the average (over all endpoints) false positive rate per case. AU-FROC can summarize performance and give an overall evaluation of both RP2 and LC prediction.

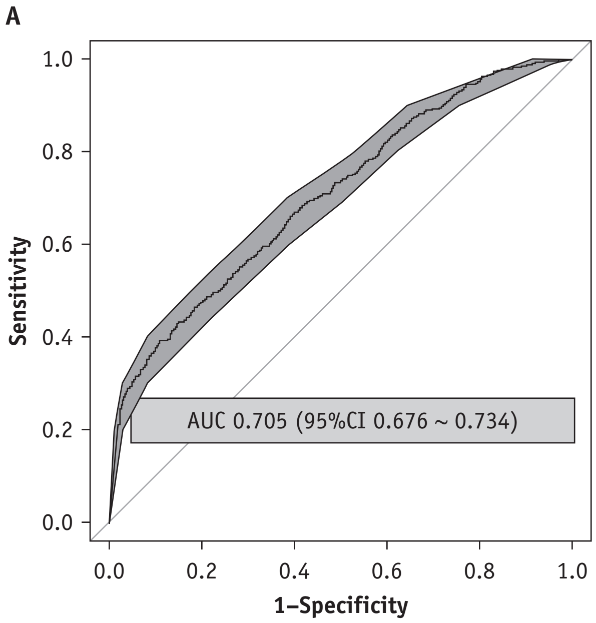
Stratified 5-fold cross-validation was conducted on the 117 patients per TRIPOD level 2 type-a criterion[27]; that is, data were randomly split into 2 groups: one for model development, the other for evaluation of performance. A total of 100 different splits were conducted to consolidate the evaluation, and 95% confidence intervals (CIs) were deduced with the quantile function of the norm distribution after the variance of the c-index was calculated as defined by DeLong.[43] Additionally, a data set containing 25 newly treated patients was used in an independent internal test following the TRIPOD level 2 type-b criterion; that is, data were split based on time, which is thought of as a better design compared with random splits. Moreover, 327 patients from the RTOG 0617 protocol were considered for external validation (TRIPOD level 3).

Results

Performance of *ADNN*-models and analytical models

Lyman/log-logistic models were trained, tested, and validated in the same way as the proposed models for comparison purposes. Their optimal values of free parameters are presented in Appendix E3**.** Cross-validated c-indexes with 95% confidence intervals for both analytical and *ADNN* models were calculated and summarized in Table 1. The corresponding ROC curves of RP2 and LC prediction by the best models *ADNN-com-joint* are presented in Figure 3. Our models were independently tested on a data set of an additional 25 prospectively treated patients at our institution following TRIPOD level 2b, and the corresponding results were shown in Table 1. For external validation (TRIPOD level 3), as RP2 was not available in the RTOG 0617, RP3 prediction was evaluated instead of RP2, which provides higher sensitivity to toxicity. A total of 327 patients were used to validate the proposed models, with results shown in Table 1.

Generally, *ADNN-DVH* models, which were based solely on dosimetric information, outperformed traditional Lyman/log-logistic models in RP2/LC prediction, in both independent internal test and external validation. *P* values for comparison of all the *ADNN models* and analytical models are shown in Table 1. Architectures *ADNN-com,* which incorporated image and biological information, further improved the performance compared with *ADNN-DVH.* Architectures of *ADNN-com-joint* models, which are based on *ADNN-com,* showed comparable results and further realized joint prediction, yielding a cross-validated c-index of 0.705 (95% CI, 0.676-0.734) on RP2 prediction, a cross-validated c-index of 0.740 (95% CI, 0.715-0.765) on LC prediction, and a cross-validated joint-prediction AU-FROC of 0.729 (95% CI, 0.697-0.773).



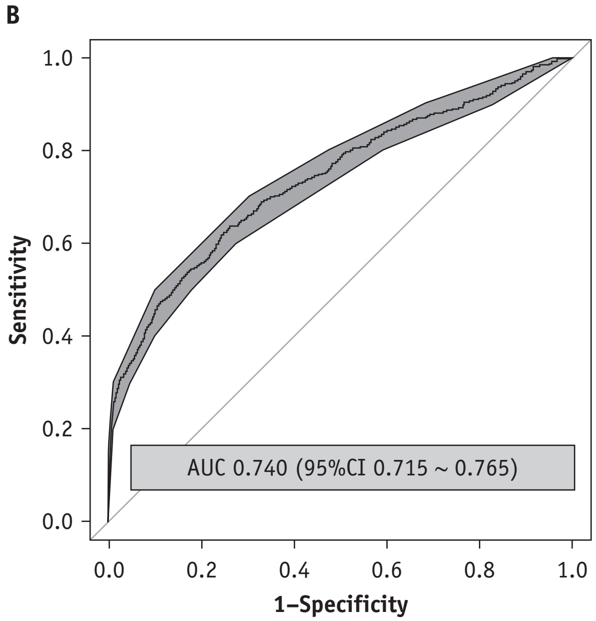


Fig. 3. Receiver operating character (ROC) curves of (A) radiation pneumonitis 2 (RP 2) and (B) local control prediction by ADNN-com. *Abbreviation*: AUC = area under ROC curve.

Activation map Grad-CAM of *ADNN* models

Based on Figure 4A-C for RP2 prediction, in the first convolutional layer, low-dose regions are highlighted, which may be attributed to the distribution in the original inputs (ie, lung DVHs). The highlighted regions then gradually shift toward higher-dose regions in the second and third convolutional layers as CNN learns that higher-dose regions are more important in discriminating toxicity. Similarly, for LC prediction as in Figure 5A-C, the highlighted region corresponds to the peak in original inputs (ie, tumor DVHs in the first convolutional layer). In the second and third convolutional layers, CNN gradually spread out highlighted regions, looking for characteristics outside the highlighted region in the lower-level layers.

In Figures 4D, 5D, and 5E, Grad-CAMs in the hidden layers in VAEs are shown. Figures 4D and 5D demonstrate the Grad-CAMs in the same layer but are based on different outputs (RP2 vs LC). The difference in Grad-CAM between toxicity groups (Fig. 4D) is more obvious than in Figure 5D, which indicates cytokines contribute more to the prediction of RP2 than to LC prediction. From Figure 5E, differences between LC = 0 and LC = 1 in Grad-CAMs are shown, which indicates that PET information contributes to discriminating LC.

Discussion

In this study, 3 deep learning architectures, *ADNN-DVH, ADNN-com,* and *ADNN-com-joint,* were proposed for actuarial prediction of RP2 and LC in NSCLC patients. Compared with models that have been previously applied to outcome prediction (eg, analytical models, support vector machine,[44] Bayesian network,[13] and NN models[45]), our proposed models can incorporate a larger and heterogeneous amount of dosimetric, image, and biological data. Second, the proposed architectures realized simultaneous dimensionality reduction and prediction, eliminating the necessity of feature selection in conventional multivariable prediction models. Third, time-to-event information was incorporated for actuarial prediction. Fourth, in *ADNN-com-joint,* multi-endpoint prediction is realized, which potentially helps take into account cross-talk between competing outcomes for clinical decision support. The performance of our models was evaluated in a complex multilevel evaluation process including independent internal tests TRIPOD level 2 type a, level 2 type b, and external validation—TRIPOD level 3.

In *ADNN-DVH* models, except for differential DVHs, volumes and mean doses were also used as inputs. Differential DVHs were chosen instead of absolute DVHs as they demonstrate more variance in morphologic characteristics among patients, which would benefit the prediction task. An experiment of using solely differential DVHs as inputs was also done, with details provided in Appendix E4. The alternative method in which DVHs were the sole inputs yielded inferior performance compared with *ADNN-DVH.* This could be attributed to the fact that in the *ADNN-DVH,* the additional inputs of volume and mean dose may force the 1D CNN to focus on learning morphologic characteristics rather than being distracted by average properties across the DVH.

To investigate the contribution of patient-specific information, including imaging and biological data, to the prediction of RP2 and LC, values of each category's patient-specific data were forced to be 0. The corresponding c-indexes calculated from architecture *ADNN-com-joint* are shown in Appendix E4. According to these results,

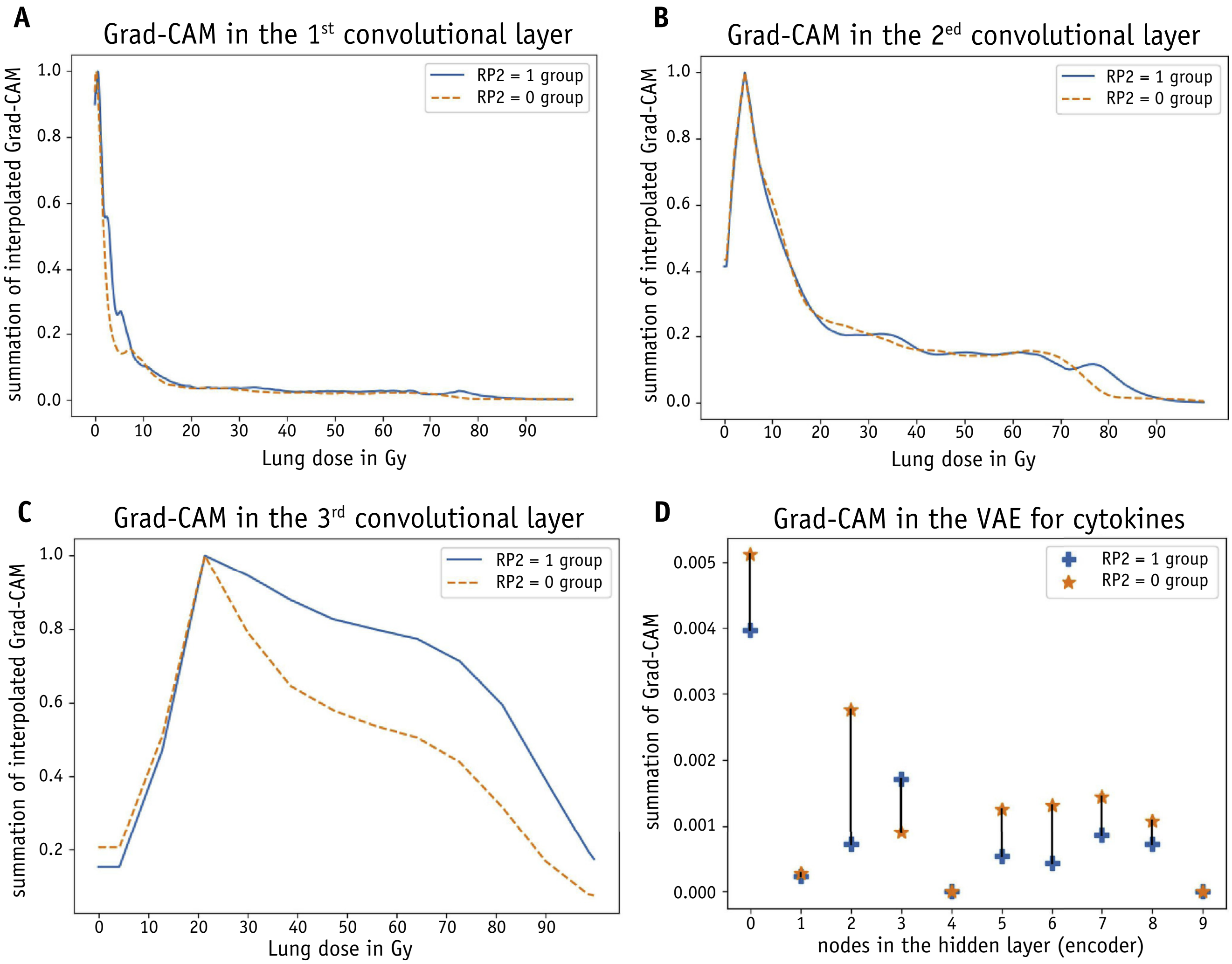


Fig. 4. Visualization by gradient-weighted class activation mapping (Grad-CAMs) in our proposed models for radiation pneumonitis 2 (RP 2) prediction. (A, B, and C) Gradient-weighted class activation mapping (Grad-CAMs) (an average for a specific group) in the 3 convolutional layers in *ADNN-DVH* models; (D) Grad-CAMs (an average for a specific groups) in the hidden layers of encoders in variational encoders for cytokines.

ADNN-com without cytokine information showed the worst performance in RP2 prediction; hence, cytokines contributed most to RP2 prediction aside from dosimetric information. Similarly, *ADNN-com-joint* without PET information as inputs showed the worst performance in LC prediction; therefore, aside from dosimetric information, PET radiomics contributed most to LC prediction. Taken together, cytokines that are related to physiological processes such as inflammation can provide valuable information[46] in addition to dosimetric information in RP2 prediction. Texture features calculated from the tumor region of functional imaging PET[29] are able to supplement dosimetric information in LC prediction. By incorporating complex radiation therapy interactions among biological, imaging, and physical data, better outcome prediction can be achieved.

In our study, Grad-CAM was applied to the convolutional layers to produce coarse localization maps highlighting important regions for predicting outcomes (ie, LC and RP2). It provides insights into interpreting how *ADNN-DVH* models work. Although the original Grad-CAM is designed for convolutional layers, a modified version of Grad-CAM for fully connected layers was also deployed. It showed its ability to understand the contribution of different patient-specific information to a specific outcome in *ADNN-com-joint* models.

Missing imaging and biological information were imputed by their median values in our study. Additionally, patients with missing information were assigned only to the training set in cross-validations to ensure the test set only contained patients with complete and accurate information. Median imputation is simple and fast. However, because it does not factor in the correlation among variables, the performance of the predictive model might be affected. More sophisticated imputation methods such as K nearest neighbors,[47] tree-based imputation,[48] and multiple imputation[49] can be also explored.

Because RP2 and LC events usually happened in the first several months after treatment (ie, time-to-event

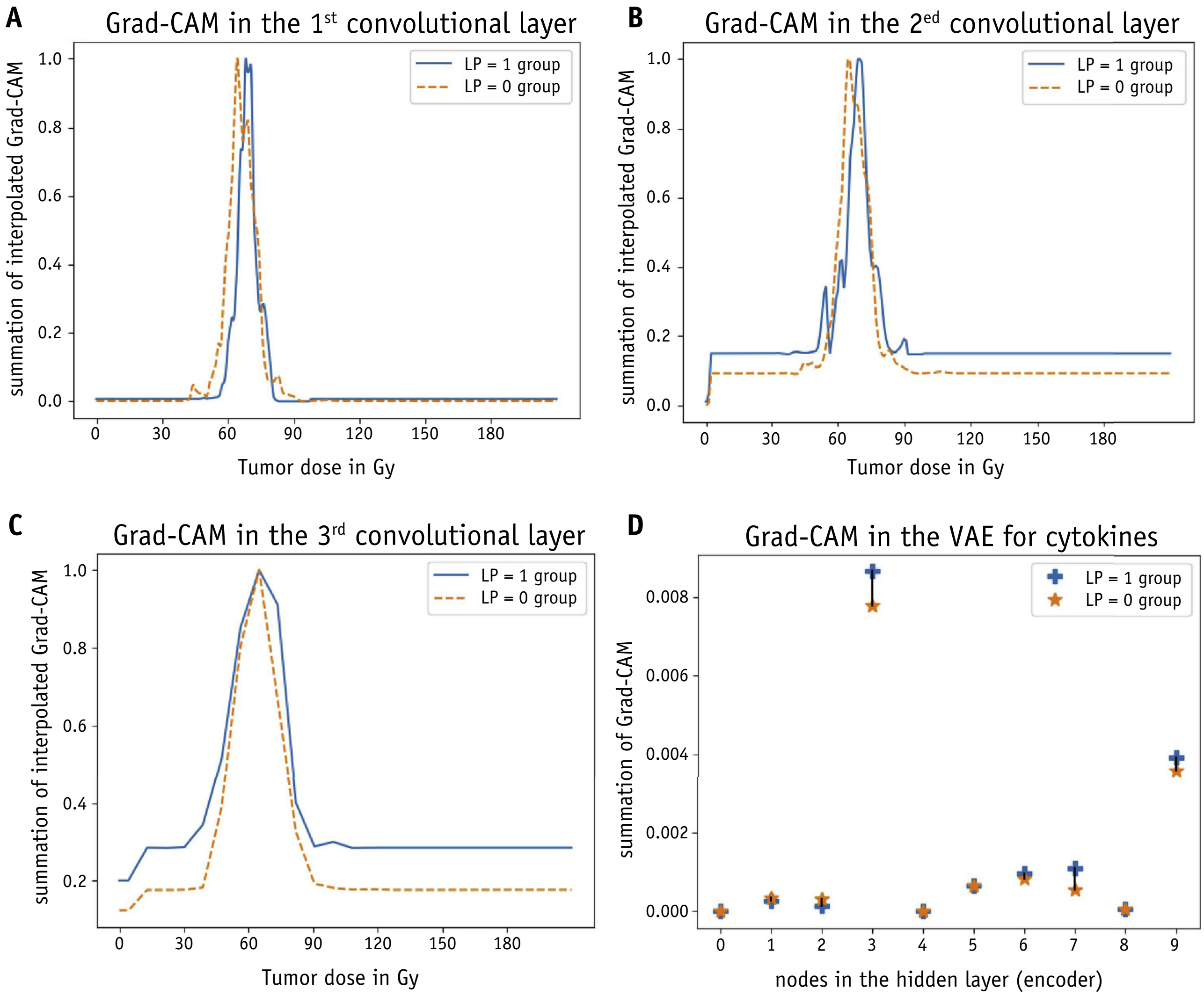


Fig. 5. Visualization by gradient-weighted class activation mapping (Grad-CAMs) in our proposed models for local control prediction. (A, B, and C) Grad-CAMs (an average for a specific group) in the 3 convolutional layers in *ADNN-DVH* models; (D and E) Grad-CAMs (an average for a specific groups) in the hidden layers of encoders in VAEs for cytokines and PET, respectively. *Abbreviations:* PET = positron emission tomography; VAE = variational encoder.

distributions are left skewed), the range of time intervals was intentionally determined such that each would contain the same number of events. As a result, parameters related to each interval would be trained equally with similar numbers of events/censor data. However, as oversampling method SMOTE[50] is applied in classification problems, oversampling methods[51] that can model time to survival may be further investigated in the actuarial analysis.

In this study, patients from TCGA data sets were used for the development of the model. A composite architecture *ADNN-miRNA* as shown in Appendix E2 was trained to jointly predict LC and overall survival based on miRNA levels. The encoder of the VAE in *ADNN-miRNA* then was applied in *ADNN-com* for extraction of miRNA's latent variables. Instead of using only an LC endpoint in TCGA data, an overall survival endpoint was also considered because it might contribute additional information. The usage of transfer learning enables more reliable and generalizable representation learning; it also helped reduce overfitting and the burden of training (ie, a smaller number of parameters needed to be trained in *ADNN-com*). However, this failed to use miRNA information in our institutional data, which may be worth investigating in the future.

In this study, PET tumor radiomics were used instead of CT tumor radiomics. There are several reasons. First, FDG-PET can better monitor the response of tumors to treatment.[52,53] High metabolic rate in primary and regional nodes is known to be associated with poor treatment response and poor tumor control.[54] Second, the FDG-PET was used to improve the accuracy of target definition (GTV) in this cohort. However, CT features may provide

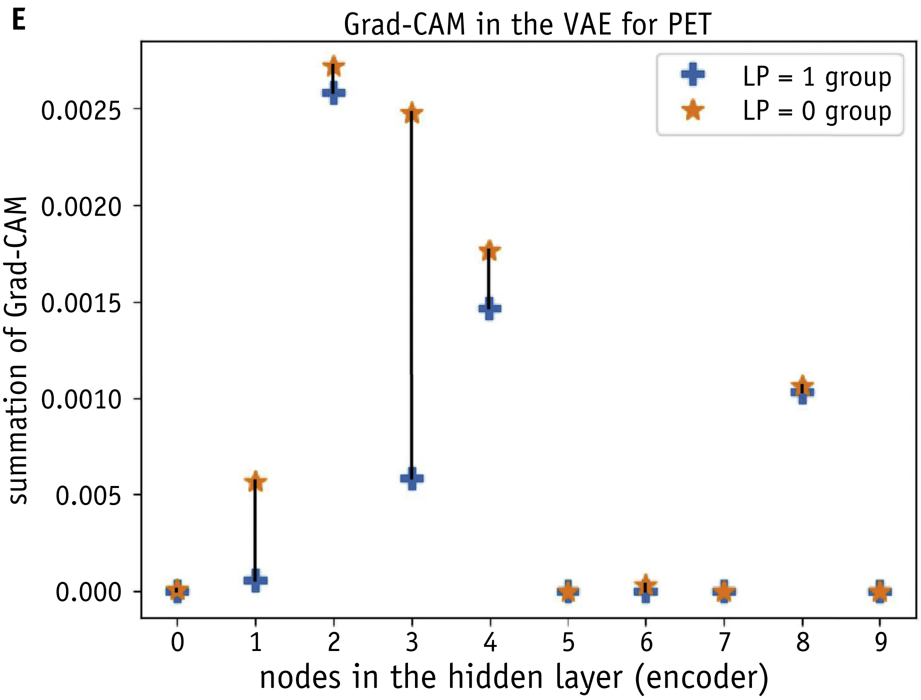


Fig. 5. (continued)

complementary information[55]; the combination of CT and PET radiomics for outcome prediction is to be explored in the future.

For external validation, *ADNN-DVH* models were evaluated for LC and RP prediction with the RTOG 0617 data set. Because RP3 rather than RP2 information was available in the RTOG 0617 data set, *ADNN-DVH* models trained for RP2 prediction on our internal data set were directly applied and analyzed for RP3 prediction under the assumption that patients who were more likely to have RP2 were also more likely to have RP3. According to our result, the c-index of RP3 is even higher than the cross-validated c-index of RP2 on our internal data set. This is expected because RP3 represents a more severe and rarer case, and physicians would have more confidence in grading RP3. Additionally, our models showed a slightly better performance in patients in arms 1 and 2 (without cetuximab) compared with patients in arms 3 and 4 (with cetuximab) with an RP3 c-index of 0.830 versus 0.684, respectively. This may indicate the role of additional factors not considered in our training set in the prediction of cetuximab-related RP toxicities in comparison with conventional concurrent chemoradiation. Further external validation of *ADNN-com* models, which are based on additional imaging and biological information, might be expected once PET imaging and biospecimens from the RTOG 0617 data set become available.

It has been shown in previous studies[56] that regional differences exist in lung radiosensitivity after radiation therapy for NSCLC patients; that is, the incidence of RP correlated significantly with the dose of different lung subvolumes (eg, posterior, caudal, ipsilateral, central, and peripheral regions). Therefore, 3D CNNs that might incorporate location information were investigated and directly applied to 3D dose distributions for the actuarial prediction of RP2 in our study. Several different architectures including 3D CNN, 3D dense Net,[57] and 3D Res-Net[58] were investigated. A model *ADNN-3Ddose* based on 3D CNN architecture, which had three 3D convolutional layers, each followed by an average pooling layer, and 2 fully connected layers, showed the best performance, as shown in Appendix E4. Note that our architecture was blinded to mean lung dose, which means relative doses at different locations provided extra information beyond the mean dose to the prediction of RP2. However, further adding mean lung dose to this architecture by concatenating it with 3D dose's reduced representation did not help improve the results. Investigation of how to better integrate such information is still ongoing.

Similarly, an *ADNN-com* architecture that can take 3D images directly as input will be of great interest. This kind of architecture can automate the whole training pipeline, potentially learning efficient features from the raw image instead of hand-crafted features. Considering the limited data size, the already complex model, and the complexity of image data (compared with 3D dose distribution), we opted to use predefined radiomics features. In the future, a fully automated pipeline is worth exploring when a larger sample size or appropriate transfer learning models become available.

Conclusions

In this study, novel deep learning architectures *ADNN-DVH, ADNN-com*, and *ADNN-joint* were proposed for the prediction of RP2 and LC in NSCLC patients. These proposed models systemically yield better performance than prevalent analytical models and generalized Lyman and log-logistic models for RP2 and LC predictions. Additionally, these architectures enabled simultaneous dimensionality reduction and prediction, eliminating the necessity of feature engineering in traditional machine learning methods. Time-to-event information was also incorporated into these models to realize actuarial prediction. Furthermore, large heterogeneous data sets, including PET tumor radiomics and biological and dosimetric information, were considered in architectures *ADNN-com* and *ADNN-com-joint. ADNN-com-joint,* which predicted multi-endpoints, showed the best performance and could potentially consider trade-offs and cross-talk between competing risk outcomes for better clinical decision support.

References

1. El Naqa I. *A Guide to Outcome Modeling in Radiotherapy and Oncology: Listening to the Data* 1st ed. Portland, USA: CRC press; 2018.
2. Friedland W, Kundrat P. Example applications in oncology. In: *A Guide to Outcome Modeling in Radiotherapy and Oncology: Listening to the Data*. 1st ed. Portland, USA: CRC press; 2018.
3. Tyldesley S, Boyd C, Schulze K, et al. Estimating the need for radiotherapy for lung cancer: An evidence-based, epidemiologic approach. *Int J Radiat Oncol Biol Phys* 2001;49:973-985.
4. Socinski MA, Rosenman JG, Halle J, et al. Dose-escalating conformal thoracic radiation therapy with induction and concurrent carboplatin/paclitaxel in unresectable stage IIIA/B nonsmall cell lung carcinoma: A modified phase I/II trial. *Cancer* 2001;92:1213-1223.
5. Kong F-M, Ten Haken RK, Schipper MJ, et al. High-dose radiation improved local tumor control and overall survival in patients with inoperable/unresectable non-small-cell lung cancer: Long-term results of a radiation dose escalation study. *Int J Radiat Oncol Biol Phys* 2005;63:324-333.
6. Li AX, Alber M, Deasy J. The use and QA of biologically related models for treatment planning: Short report of the TG-166 of the therapy physics committee of the AAPM. *Med Phys* 2012;39:1386-1409.
7. Lyman JT. Complication probability as assessed from dose-volume histograms. *Radiat Res Suppl* 1985;8:S13-S19.
8. Cui S, Haken RKT, Naqa IE. Building a Predictive Model of Toxicity: Methods. In: Rancati T, Fiorino C, eds. Modelling Radiotherapy Side Effects: Practical Applications for Planning Optimisation. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2019.
9. Qi XS, Guerrero M, Li X. Outcome Modeling in Treatment Planning. In: Naqa IE, ed. A guide to Outcome modeling in radiotherapy and oncology: Listening to the Data, 2018. Portland, USA: CRC press; 2018.
10. El Naqa I. Modeling of tumor control probability (TCP). In: Machine Learning in Radiation Oncology: Theory and Applications. New York, NY: Springer International Publishing; 2015.
11. Velec M, Haddad CR, Craig T. Predictors of liver toxicity following stereotactic body radiation therapy for hepatocellular carcinoma. *Int J Radiat Oncol Biol Phys* 2017;97:939-946.
12. El Naqa I, Bradley JD, Lindsay PE, et al. Predicting radiotherapy outcomes using statistical learning techniques. *Phys Med Biol* 2009; 54:S9-S30.
13. Luo Y, McShan DL, Matuszak MM. A multiobjective Bayesian networks approach for joint prediction of tumor local control and radiation pneumonitis in nonsmall-cell lung cancer (NSCLC) for response-adapted radiotherapy. *Med Phys* 2018;45:3980-3995.
14. Luo Y, Tseng H-H, Cui S, et al. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR Open* 2019;1:20190021.
15. Boldrini L, Bibault J-E, Masciocchi C, et al. Deep learning: A review for the radiation oncologist. *Front Oncol* 2019;9:977.
16. Cui S, Tseng H-H, Pakela J, et al. Introduction to machine and deep learning for medical physicists. *Med Phys* 2020;47:e127-e147.
17. Luo Y, Chen S, Valdes G. Machine learning for radiation outcome modeling and prediction. *Med Phys* 2020;47:e178-e184.
18. Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. *Front Oncol* 2020;10:790.
19. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: Data, validation, communication challenges. *Med Phys* 2018;45:e834-e840.
20. Cui S, Luo Y, Tseng H-H, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Med Phys* 2019;46: 2497-2511.
21. Cui S, Luo Y, Tseng H-H, Haken RKT, Naqa IE. Artificial neural network with composite architectures for prediction of local control in radiotherapy. *IEEE Trans Radiat Plasma Medl Sci* 2019;3:242-249.
22. Kigma D, Velling M. Auto-encoding variational bayes, 2014.
23. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ* 2019;7:e6257.
24. Tucker SL, Liu HH, Liao Z. Analysis of radiation pneumonitis risk using a generalized Lyman model. *Int J Radiat Oncol Biol Phys* 2008; 72:568-574.
25. Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol* 2018;14:e1006076.
26. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B* 1972;34:187-220.
27. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;162:55-63.
28. Bradley JF. Data from NSCLC-Cetuximab. The Cancer Imaging Archive. Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=33948334. Accessed March 1, 2021.
29. Vallières M, Freeman C, Skamene S, et al. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol* 2015;60:5471-5496.
30. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295:328-338.
31. Clark K, Vendt B, Smith K. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging* 2013;26:1045-1057.
32. Albertina B, Watson M, Holback C. Radiology data from the Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection. *Cancer Imaging Archive.* Published online 2016. Available at: https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD. Accessed March 1, 2021.
33. Kirk S, Lee Y, Kumar P. Radiology data from the cancer genome atlas lung squamous cell carcinoma [TCGA-LUSC] collection. *Cancer Imaging Archive.* Published online 2016. Available at: https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUSC. Accessed March 1, 2020.
34. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010;22:1345-1359.

35. Bradley JD, Paulus R, Komaki R. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): A randomised, two-by-two factorial phase 3 study. *Lancet Oncol* 2015;16:187-199.
36. Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017, IEEE. Available at: https://arxiv.org/abs/1610.02391. Accessed March 1, 2021.
37. bbmle: Tools for General Maximum Likelihood Estimation program; 2020. Available at: https://cran.r-project.org/web/packages/bbmle/index.html. Accessed March 1, 2020.
38. AaG P, Sam, Massa F, et al. An imperative style, high-performance deep learning library. HWaHLaABaFdaEFaR G, ed. *Advances in Neural Information Processing Systems 32: Curran Associates, Inc.* 2019(8024). Available at: https://arxiv.org/abs/1912.01703. Accessed March 1, 2021.
39. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization; 2014. Available at: https://arxiv.org/abs/1412.6980. Accessed March 1, 2021.
40. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30:1105-1117.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29-36.
42. 5. Extensions to Conventional ROC Methodology: LROC, FROC, and AFROC. *J ICRU* 2008;8:31-35.
43. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988;44:837-845.
44. Klement RJ, Allgäuer M, Appold S, et al. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* 2014;88:732-738.
45. Cui S, Luo Y, Tseng H-H, et al. Artificial neural network with composite architectures for prediction of local control in radiotherapy. *IEEE Trans Radiat Plasma Med Sci* 2019;3:242-249.
46. Kainthola A, Haritwal T, Tiwari M. Immunological aspect of radiation-induced pneumonitis, current treatment strategies, and future prospects. *Front Immunol* 2017;8:506-506.
47. Zhang S. Nearest neighbor selection for iteratively kNN imputation. *J Syst Softw* 2012;85:2541-2552.
48. Rahman MG, Islam MZ. Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowl Based Syst* 2013;53:51-65.
49. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: An overview and case study. *Emerg Themes Epidemiol* 2017;14:8.
50. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013;14:106.
51. Chapfuwa P, Tao C, Li C. Adversarial Time-to-Event Modeling, 2018. Available at: https://arxiv.org/abs/1804.03184. Accessed March 1, 2021.
52. Choi NC, Fischman AJ, Niemierko A, et al. Dose-response relationship between probability of pathologic tumor control and glucose metabolic rate measured with FDG PET after preoperative chemoradiotherapy in locally advanced non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys* 2002;54:1024-1035.
53. Hicks RJ, Mac Manus MP, Matthews JP, et al. Early FDG-PET imaging after radical radiotherapy for non—small-cell lung cancer: Inflammatory changes in normal tissues correlate with tumor response and do not confound therapeutic response evaluation. *Int J Radiat Oncol Biol Phys* 2004;60:412-418.
54. Jeong H-J, Min J-J, Park JM, et al. Determination of the prognostic value of [(18)F]fluorodeoxyglucose uptake by using positron emission tomography in patients with non-small cell lung cancer. *Nucl Med Commun* 2002;23:865-870.
55. Vaidya M, Creach KM, Frye J, et al. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiother Oncol* 2012;102:239-245.
56. Seppenwoolde Y, De Jaeger K, Boersma LJ, et al. Regional differences in lung radiosensitivity after radiotherapy for non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys* 2004;60:748-758.
57. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017:2261-2269.
58. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV, USA, 2016, pp. 770-778, https://doi.org/10.1109/CVPR.2016.90.