radiation between African-American (AA) men compared to non-Hispanic White (NHW) men (Spratt et al. [abstract ASTRO 2018]). The basis of this racial difference in radiosensitivity is likely based on variations in the expression of genes encoding for radiation response pathways. For example, decreased double strand break repair gene expression is associated increased radiosensitivity in somatic and cancer cell lines. We hypothesized that a racial difference in the expression levels of genes participating in radiation response pathways could be identified via a machine learning approach.

Materials/Methods: We extracted the gene expression level data of 7,470 patients from the Genomic Data Commons Pan-Cancer database who had race identified as AA ($n=802$) or NHW ($n=6,668$). For each patient, the expression levels of 741 genes that are known to be involved in radiation response pathways were selected for subsequent analysis. An ensemble of five machine learning methods (support vector machine (SVM), linear discriminant analysis (LDA), gradient boosted machine (GBM), Bayesian generalized linear model (BGLM), and sample mean (SM)) was trained on 80% of the data to predict for race based on this 741-gene expression panel. Out-of-sample error was estimated using 5-fold cross validation. The trained ensemble model was used to predict on the remaining 20% of the data. Performance of the ensemble model was evaluated via area under the curve (AUC) of the receiver operating characteristic curve.

Results: The mean squared error for the SVM, LDA, GBM, BGLM, and SM methods were 0.071, 0.075, 0.081, 0.076, and 0.096 respectively. The ensemble model achieved a mean square error of 0.068. Prediction by the ensemble model yielded an AUC of 0.861 (95% CI 0.844-0.878).

Conclusion: Expression levels of radiation response pathway genes can be used to accurately identify race via an ensemble of machine learning models. This supports the emerging evidence that race may be associated with radiosensitivity via intrinsic biologic differences in gene expression levels. Further studies are warranted to investigate whether these gene expression differences translate to clinically detectable variation in radiosensitivity and tumor control among different patient populations.

Author Disclosure: **R. van Dams**: None. **A.U. Kishan**: None. **N.G. Nickols**: None. **A. Raldow**: Consultant; Intelligent Automation, Inc. **C.R. King**: None. **A.J. Chang**: None. **P.A. Kupelian**: None. **M.L. Steinberg**: None. **C. Wang**: None.

2315

A Machine-Learning Model Using Artificial Neural Network to Facilitate Liver SBRT Prescription Selection By Predicting Normal Liver Geud Based on Geometric Properties of Liver and PTV

Y. Wang; *Massachusetts General Hospital, Harvard Medical School, Boston, MA*

Purpose/Objective(s): Normal liver dose is of paramount in liver SBRT due to the risk of radiation induced liver disease (RILD). In our clinic, the generalized equivalent uniform dose (gEUD) of normal liver (liver-GTV) is used to determine if a prescription (Rx) needs to be deescalated (e.g., from 50 to 45 Gy) to limit the risk of RILD. To estimate the normal liver gEUD, a planner often needs to create a preliminary plan with full Rx. The goal of this work was to build a machine learning model using artificial neural network trained with our prior SBRT patients to predict normal liver gEUD purely using the geometric properties of liver and PTV, to eliminate the need for the preliminary plan.

Materials/Methods: The initial model was trained by 40 consecutive liver SBRT patients treated in our clinic from December 2014 to February 2018 meeting the following criteria: (1) 50 Gy in 5 or 6 fractions, (2) single-lesion IMRT or VMAT plan created using multi-criteria optimization (MCO), (3) no prior radiation or no dosimetric impact from any prior radiation, and (4) 98% of GTV covered by >50 Gy. An A value of 0.9 was used to calculate the gEUD of normal liver (excluding GTV). The artificial neural network consisted of five layers including three hidden ones. It used six inputs: the x, y, and z coordinate of the center of PTV relative to the center of total liver (including GTV), the distance between the two centers, and the volume of the PTV and total liver. The model was tested on new consecutive patients since February 2018. Since the model is bounded to input ranges, a new patient with input outside the initial range cannot be used test the model. For this reason, five new patients were added to the learning data before ten qualified new patients were found to test the model. All training and test patients were treated on two different linear accelerators, both with 5-mm leaves in central 20×20 cm^2 field.

Results: The model accurately predicted normal liver gEUD for five patients, with an error of $<3.2\%$ (<0.35 Gy). For the other five, the error ranged from 7.9% to 12.9% (0.76 to 1.33 Gy). On average, the error was $-0.8\pm8.4\%$ (-0.14 ± 0.98 Gy). The error increased when any input approached the limit in the under-sampled boundary region. For the patient with the largest error, the x of PTV center was -7.23 cm, only 0.08 cm away from the lower limit. For the patient with the second largest error, the y of PTV center was 7.16 cm, only 0.75 cm away from the upper limit. The model will become more robust once it learns from those patients with inputs in the boundary regions. The calculation for a new patient takes <1 second on a regular PC. The model can be exported as a Python function and scripted in common contouring and planning systems.

Conclusion: A machine learning model was built to predict normal liver gEUD in 50-Gy liver SBRT, only using the geometric properties of liver and PTV. The prediction was within 13% (1.3 Gy) of the actual value for ten consecutive prospective patients. The model can facilitate the Rx selection in liver SBRT, promoting efficiency for planners and clinicians.

Author Disclosure: **Y. Wang**: None.

2316

Machine Learning to Predict Toxicity in Head and Neck Cancer Patients Treated with Definitive Chemoradiation

A.P. Wojcieszynski, Jr,[1] W. La Cava,[2] B.C. Baumann,[3] J.N. Lukens,[4] A. Fotouhi Ghiam,[5] R.J. Urbanowicz,[2] S.D. Swisher-McClure,[5] A. Doucette,[5] P.E. Gabriel,[5] A. Lin,[5] Y. Xiao,[1] J.H. Moore,[2] and J.M. Metz[5]; [1]*University of Pennsylvania, Philadelphia, PA,* [2]*Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA,* [3]*Washington University School of Medicine, Department of Radiation Oncology, St. Louis, MO,* [4]*Department of Radiation Oncology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA,* [5]*Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA*

Purpose/Objective(s): Concurrent chemoradiation (CRT) is one of the standard-of-care curative treatments for patients with head and neck cancer, but is associated with substantial morbidity. Proton radiation therapy (RT) may allow for decreased toxicity due to treatment related side-effects, but prospective studies are limited. We set out to use a machine learning (ML) approach to determine factors associated with morbidity in patients treated with CRT for head and neck cancer.

Materials/Methods: Head and neck cancer patients treated with definitive CRT from 2011-2016 were identified. An IRB-approved review was performed, with CTCAE toxicity prospectively collected during follow-up. A regression and machine learning analysis of 90 and 180-day grade 3 or higher toxicity as a function of 46 patient covariates was performed. Data was cleaned using pairwise correlation filtering with a Pearson's correlation of 0.6 to reduce collinearity. A multivariate logistic regression model was trained to determine significant factors associated with toxicity. Three ML methods were utilized for predicting toxicity: penalized logistic regression, random forest, and XGBoost, a gradient boosting method. For each method, we conducted 10-fold cross validation on a training subset of the data to tune each method's hyperparameters. The tuned models were assessed according to their predictions on a hold-out test set. We repeated this analysis with 30 random shuffles of the data to generate robust performance estimates. In addition, we collated feature importance measures from each final model to interpret the importance of each covariate. We compared the important factors for prediction to those found to be associative on regression.

Results: 437 patients were included in the analysis, 397 treated with photon RT, and 40 treated with proton RT. Patient characteristics were well balanced between the cohorts, with no difference in age, sex, stage, RT dose, or chemotherapy type. Using regression, increased integral radiation dose to regions outside of the PTV was associated with increased Grade 3+ toxicity at both 90 and 180 days ($p = 0.03$ and $p = 0.02$), and the use of proton RT trended strongly ($p = 0.07$ and $p = 0.04$) with decreased toxicity. Other marginal effects were observed for insurance provider ($p = 0.05$) and attending physician ($p = 0.05$). Using ML, we were able to predict toxicity with moderate success for the 90-day (c-statistic: 0.65) and 180-day (c-statistic: 0.63) observations with the random forest approach. We found good agreement between the most important features for the ensemble tree methods, which were, in order of decreasing importance, PTV integral dose, body mass index, integral dose to regions outside the PTV, and age.

Conclusion: Using a ML approach, we were able to grade 3 toxicity in patients undergoing CRT for head and neck cancer with moderate success. PTV integral dose and integral dose to regions outside of the PTV were associated with increased toxicity, and may support the use of proton RT in this population.

Author Disclosure: **A.P. Wojcieszynski**: None. **W. La Cava**: None. **B.C. Baumann**: Employee; University of Pennsylvania School of Medicine. **J.N. Lukens**: None. **A. Fotouhi Ghiam**: None. **R.J. Urbanowicz**: None. **S.D. Swisher-McClure**: None. **A. Doucette**: None. **P.E. Gabriel**: None. **A. Lin**: Employee; Children's Hospital of Philadelphia. Advisory Board; Galera Pharmaceuticals. **Y. Xiao**: None. **J.H. Moore**: None. **J.M. Metz**: None.

2317

Validation of Deep Learning-based Auto-Segmentation for Organs at Risk and Gross Tumor Volumes in Lung Stereotactic Body Radiotherapy

J. Wong,[1] V. Huang,[2] J.A. Giambattista,[3] T. Teke,[4] and S. Atrchian[4]; *[1]BC Cancer, Vancouver, BC, Canada, [2]BC Cancer, Surrey, BC, Canada, [3]Saskatchewan Cancer Agency, Regina, SK, Canada, [4]BC Cancer, Kelowna, BC, Canada*

Purpose/Objective(s): Accurate contouring of organs at risk (OAR) and gross tumor volumes (GTV) is particularly important in stereotactic body radiotherapy (SBRT) where smaller margins are used. Manual segmentation is labor intensive and can suffer from significant inter-observer variability. Here we evaluate the performance of deep learning auto-segmentation models trained from retrospective manually drawn contours from a single center and assess whether these models can accurately segment patient planning CT scans from a different cancer center with acceptable results.

Materials/Methods: Auto-segmentation models were trained using a deep convolutional neural network based on a U-net architecture using 210 planning CT scans, which included 160 publicly available planning CT scans with ground truth contours reviewed by a radiation oncologist and 50 lung SBRT CT scans from a single center (center A). Deep learning models were then used to segment 100 planning CT scans, which consisted of 50 additional scans from center A and 50 planning CT scans from a separate cancer center (center B). The original clinical contours (CC) were compared with the deep learning based contours (DC) using the Dice Similarity Coefficient (DSC) and 95% Hausdorff distance transforms (DT).

Results: Comparing DCs to CCs for all 100 contoured planning CT scans, the mean DSC and 95% DT were 0.93 and 2.8 mm for aorta ($n=81$), 0.81 and 3.3 mm for esophagus ($n=99$), 0.95 and 5.1 mm for heart ($n=100$), 0.98 and 3.1 mm for lung ($n=190$), 0.56 and 6.6 mm for brachial plexus ($n=101$), 0.82 and 4.2 mm for proximal bronchial tree ($n=100$), 0.90 and 1.6 mm for spinal cord ($n=87$), 0.91 and 2.3 mm for trachea ($n=100$), and 0.71 and 5.2 mm for lung GTVs ($n=85$). The DSC and 95% DT were not significantly different for center A and center B for aorta, lung GTV, heart, lung, brachial plexus, spinal cord, and trachea. Structures with significantly different DSC or 95% DT between the two centers included the esophagus DSC (0.80 vs 0.83, $p=0.02$) and proximal bronchial tree 95% DT (3.6 vs 4.8 mm, $p=0.001$).

Conclusion: Deep-learning auto-segmentation models can provide accurate segmentation for OARs used in lung SBRT. Models trained with a single institution's data were accurate when validated on a separate institution's planning CT scans, despite variations in scan quality and contouring practices. Deep learning lung GTV segmentation models reliably located the target lesions but generally were less accurate than the organs at risk models due to the variable location and size of lung tumors. Deep learning auto-segmentation can provide an accurate starting point for review and manual adjustment and should improve efficiency in lung SBRT planning.

Author Disclosure: **J. Wong**: None. **V. Huang**: None. **J.A. Giambattista**: Co-founder; Limbus AI. **T. Teke**: None. **S. Atrchian**: None.

2318

A Radiomic Classifier of Locoregional Failure after Definitive Radiation in Head and Neck Squamous Cell Carcinoma

Y. Yuan,[1] Z. Zhang,[1] X. Pan,[2] X. Qi,[1] and R.K. Chin[1]; *[1]Department of Radiation Oncology, University of California, Los Angeles, Los Angeles, CA, [2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an, China*

Purpose/Objective(s): Biomarkers that can stratify head and neck squamous cell carcinoma (HNSCC) patients by their risk for locoregional failure (LRF) after radiation are lacking. Quantitative imaging features, or radiomics, can find imaging characteristics that are prognostic for outcomes. Therefore, we hypothesize that supervised machine learning can derive a subset of radiomic features that can predict LRF in HNSCC patients who receive definitive radiation.

Materials/Methods: Patients with HNSCC who experienced a biopsy-proven LRF after definitive radiation or chemoradiation and patients who did not experience LRF were identified in the electronic medical record. DICOM images of treatment plans for each patient were extracted with associated contours delineating the primary gross tumor volume (GTV). Radiomic features were extracted utilizing the GTV as the region of interest via an in-house algorithm. The full dataset of radiomic features was scaled and split into training and test sets. Feature selection was carried out using a randomized logistic regression model. A random forest machine learning algorithm was then trained on a reduced training set containing the most relevant features and with LRF as the outcome of interest. The trained classifier was evaluated on the test set with the area under the curve of the receiver operating characteristic (AUC) as the evaluation metric. The relapse free survival (RFS) of patients stratified by clinical staging was compared to stratification by the trained radiomics classifier using Kaplan-Meier Analysis with a log-rank test ($\alpha = 0.05$).

Results: A total of 47 patients were included in the study of whom 15 patients experienced a LRF after radiation. From the GTV of each patient a total of 1769 radiomic features were extracted corresponding to several imaging characteristic domains including shape and intensity measures, histogram of oriented gradients, gray level co-occurrence matrices (GLCM), and neighbor intensity differences. After feature selection, the feature space was reduced to the 41 most relevant radiomic features. A random forest algorithm trained on 60% of the complete dataset achieved a mean AUC of 87.5% with a standard deviation of 11.1% when evaluated on the independent test set representing 40% of the full dataset. The average relative importance of each radiomic feature in predicting for LRF was also ranked. The most predictive radiomic features were intensity and GLCM based imaging characteristics. RFS of patients stratified by the radiomic classifier was significantly different ($p = 0.0013$) while stratification by clinical stage was not ($p=0.26$).

Conclusion: Utilizing a random forest supervised machine learning algorithm trained on radiomic features extracted from the GTV contour on the planning CT we have developed a classifier of LRF after definitive radiation or chemoradiation for HNSCC. Ranking the relative contributions of the radiomic features to the model reveals imaging characteristics that are most predictive of LRF.

Author Disclosure: **Y. Yuan**: None. **Z. Zhang**: student; UCLA. **X. Pan**: None. **X. Qi**: None. **R.K. Chin**: None.