

Radiomics for Clinical Outcome Prediction in Head and Neck Cancer Patients

Tanner Street: Clemson University

Vineet Narayanan: University of New Mexico

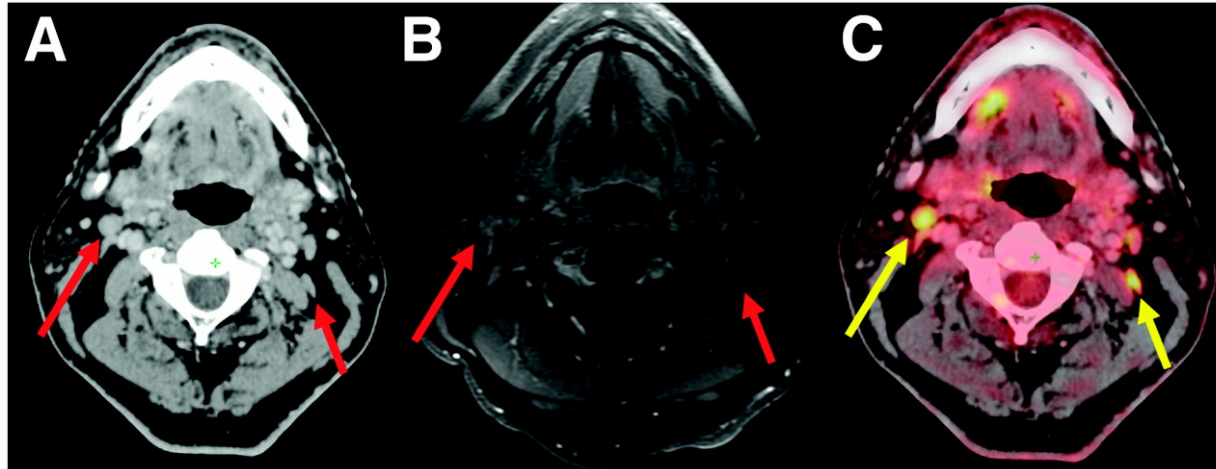
Iowa Summer Institute in Biostatistics - University of Iowa

Department of Biostatistics

Dr. Brian Smith, Professor, Dept. of Biostatistics, University of Iowa

Introduction

- **Radiomics:** Extraction and analysis of quantitative data from medical images
 - *Qualitative* analysis historically
 - Quantitative analysis → benefits



Background

- Clinical Trial Background
 - 287 participants w/ head and neck cancer
 - Goal: Evaluate diagnostic performance of PET/CT imaging
- Data from Clinical Trial
 - Clinical information
 - Quantitative imaging biomarkers
 - Clinical outcomes
- **Our Goal: Develop and evaluate statistical models to predict clinical outcomes of patients with head and neck cancer**

Data Overview

- **Clinical Outcome**

- Response variable
- Time: days until death (event) or censor
 - Median: 752 days
- Status: event or censor
 - 36 events / 181 patients

- **Quantitative Imaging Biomarkers**

- Four data sets: primary liver, total liver, primary, and total
 - Primary = primary lesion | Total = all lesions
 - Liver: normalization of data
 - Each data set tested separately
- 111 predictor variables in each data set

- **Clinical Information**

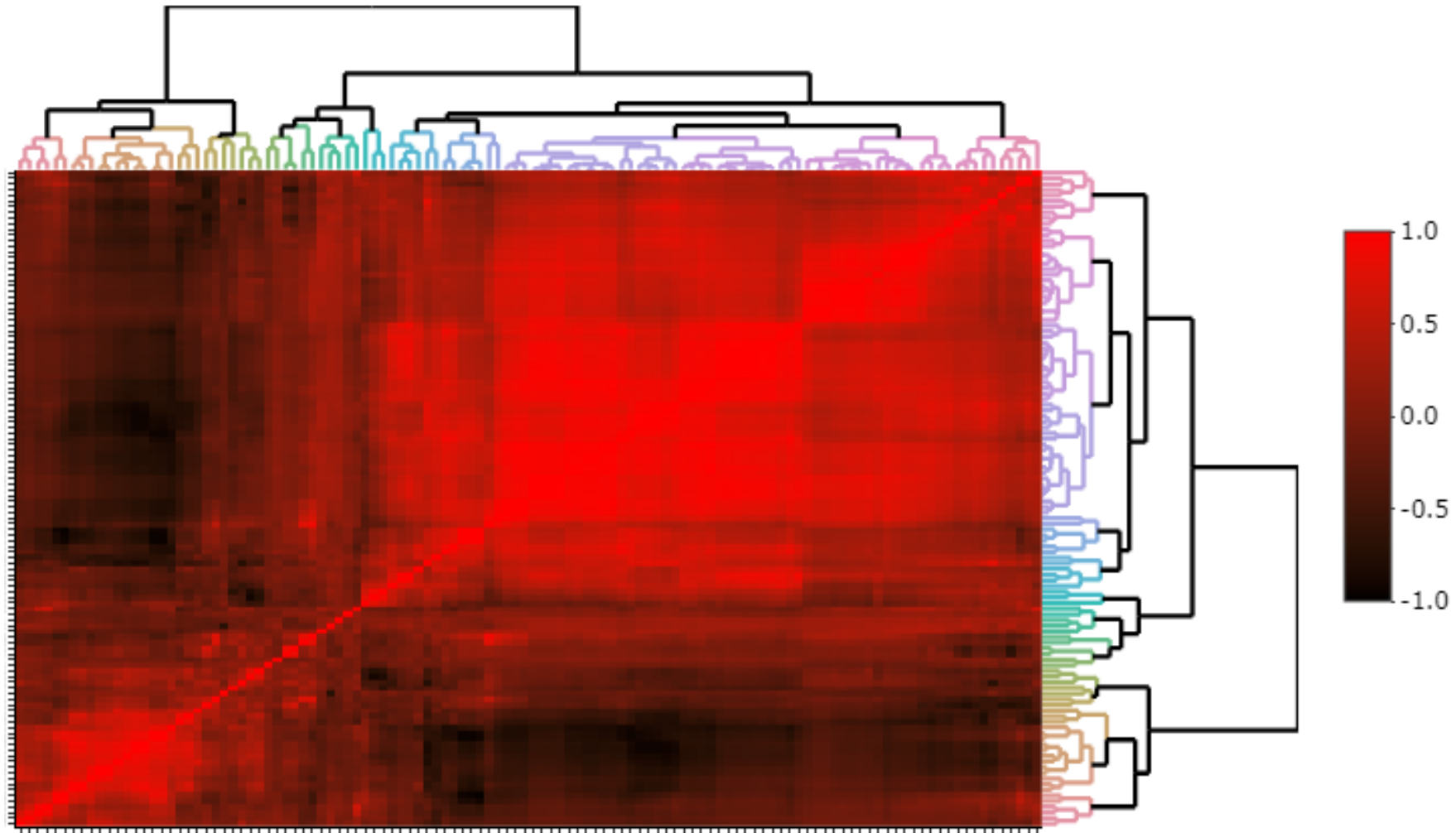
- 5 predictor variables
- Summary statistics...

Clinical Summary Statistics

	Overall (N=181)
Age	
- Mean (SD)	59.8 (11.3)
- Range	24.0 - 90.0
Sex	
- Female	65 (35.9%)
- Male	116 (64.1%)
Tumor Size	
- "Large"	69 (38.1%)
- "Small"	112 (61.9%)
Lymph Node Presence	
- Negative	153 (84.5%)
- Positive	28 (15.5%)
Race	
- Not White	44 (24.3%)
- White	137 (75.7%)

Correlation Heatmap

Clinical (5 variables) and Total Liver (111 variables)



Testing Overview

- Pre-Processing
 - K-Medoids
 - High correlation filtering
 - No filtering
- Four Models
 - Gradient Boosting Regression
 - 2 Random Forests
 - Penalized Cox regression
- Parameter tuning
- Metric: Concordance Index
 - Based on ordering, not magnitude of prediction
 - Magnitudes
 - = 0.5 → entirely random
 - = 1.0 → perfect concordance

Cross-Validation

Nested 5-Fold Cross-Validation



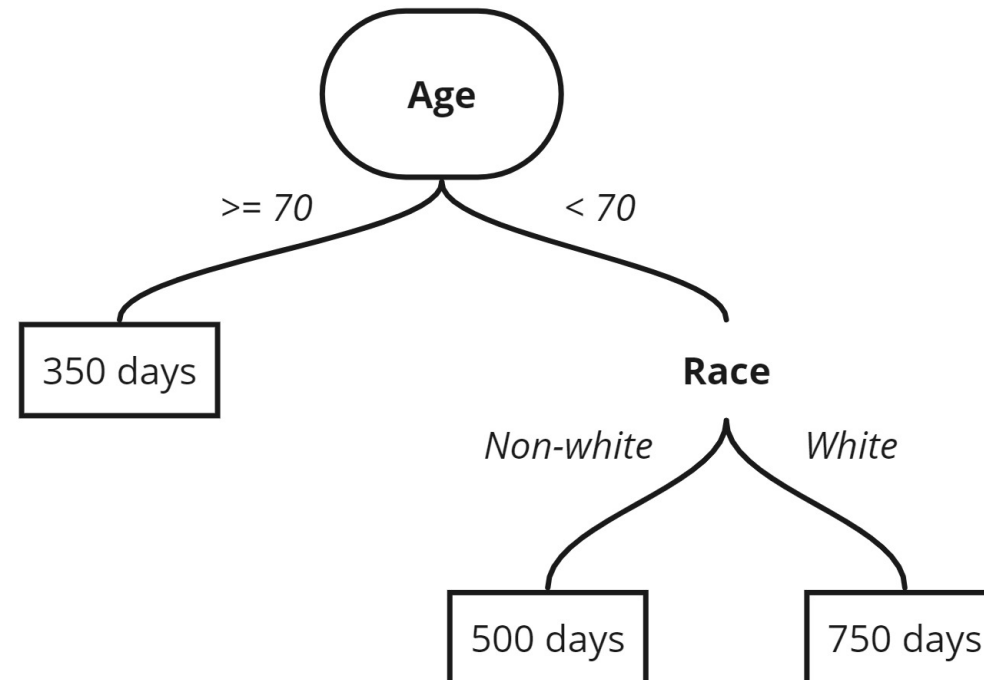
Pre-Processing Techniques

- Variable filtering was applied in pre-processing steps prior to fitting the models
- Filtering is a method of selecting a subset of variables from the original data set
- K-Medoids
 - Partitions variables into specified number of k clusters
 - Returns the cluster medoid variables as the filtered subset
 - Medoid is variable closest to cluster center
- High Correlation Filtering
 - Iteratively removes variables whose pairwise correlation magnitude are above a specified threshold
 - 47 variables selected from total liver data set with a threshold = 0.8

$$c = \sum_{k=1}^K \sum_{C(i)=k} d(x_i, m_k),$$

Tree-Based Models

- Gradient Boosting Regression and 2 Random Forests
- Models based on decision trees
- Node selection
 - Variable and threshold that minimizes some test statistic
 - Specified number of random candidates tested at each node



Tree-Based Models cont.

- Random Forests
 - Many trees, average of performance
 - RFSRC
 - 2 tuning parameters
 - Ranger
 - 1 tuning parameter
- Gradient Boosting Regression
 - Trees predict pseudo-residuals
 - 2 tuning parameters
- Non-linear variables and interactions

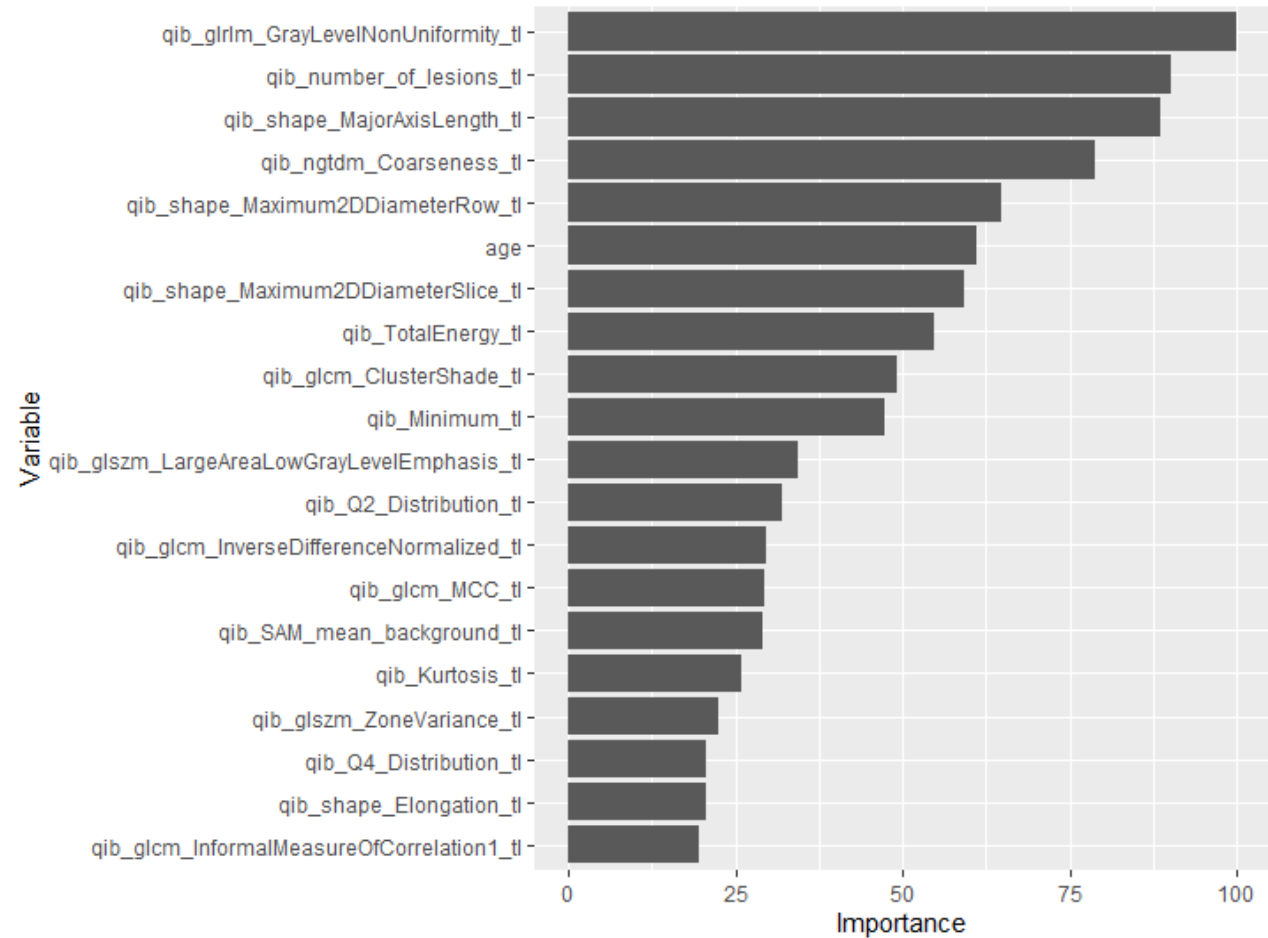
Penalized Cox Regression

- Semi-parametric regression model for censored time-to-event outcomes like overall survival
- Predictors were modeled as linear effects without interactions
- Regularize data using the elastic net penalty
- Tuning parameters
 - Alpha: controls sizes of effects of correlated predictors
 - Lambda: controls number of predictors in model

Results

	DataSet	Model	Filtering	<u>C-Index</u> Mean	SD
Tree-Based Models	Primary Liver	GBM	None	0.774	0.040
	**** Total Liver	RFSRC	High Filter	0.780	0.053
	Primary	RFSRC	KMedoids	0.753	0.049
	Total	RFSRC	KMedoids	0.756	0.044
Penalized Cox Regression	Primary Liver	GLMNet	High Filter	0.689	0.094
	Total Liver	GLMNet	None	0.705	0.142
	Primary	GLMNet	High Filter	0.689	0.114
	Total	GLMNet	High Filter	0.695	0.128
Average of all models' performance using only the clinical data	Clinical	Average	None	0.575	0.074

Variable Importance (Top 20)



Conclusion

- Addition of radiomic features improves predictive performance
- Tree-based models more effective at predicting survival outcomes
- Future work
 - Evaluate how well predictive means are calibrated
 - Evaluate predictive performance of survival probabilities at points in time
 - Explore parametric distributions for the survival outcomes

Acknowledgements

- **Dr. Brian Smith**, Professor, Dept. of Biostatistics, University of Iowa
- National Heart Lung and Blood Institute (NHLBI), grant # HL-147231.

References

- Lowe VJ, Duan F, Subramaniam RM, Sicks JD, Romanoff J, Bartel T, Yu JQM, Nussenbaum B, Richmon J, Arnold CD, Cognetti D, Stack BC Jr. Multicenter Trial of [18F]fluorodeoxyglucose Positron Emission Tomography/Computed Tomography Staging of Head and Neck Cancer and Negative Predictive Value and Surgical Impact in the N0 Neck: Results From ACRIN 6685. *J Clin Oncol*. 2019 Jul 10;37(20):1704-1712. doi: 10.1200/JCO.18.01182. Epub 2019 Feb 15. PMID: 30768363; PMCID: PMC6638599.
- Quon A, Fischbein NJ, McDougall IR, Le QT, Loo BW Jr, Pinto H, Kaplan MJ. Clinical role of 18F-FDG PET/CT in the management of squamous cell carcinoma of the head and neck and thyroid carcinoma. *J Nucl Med*. 2007 Jan;48 Suppl 1:58S-67S. PMID: 17204721.
- Smith BJ (2021). `_MachineShop: Machine Learning Models and Tools_`. R package version 2.9.0, <URL: <https://cran.r-project.org/package=MachineShop>>.

References (cont.)

- *Cox Regression (Proportional Hazards, Hazard Ratio) - StatsDirect.* (2020). Stats Direct. https://www.statsdirect.com/help/survival_analysis/cox_regression.htm
- *Generalized Linear Model (GLM) — H2O 3.32.1.4 documentation.* (2021). H2O AI. <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html#regularization>

Questions?

Parameters of Best Models (Backup Slide)

- Primary Liver: GBM w/ no filtering
 - Number of trees = 200 → (Number of trees)
 - Interaction depth = 3 → (Regulates size of tree)
- Total Liver: RFSRC w/ high correlation filter
 - Mtry = 2 → (Number of variables randomly selected for testing at each node)
 - Node size = 10 → (Minimum size of leaves)
 - Threshold = 0.8 → (Correlation)
- Primary: RFSRC w/ k-medoids
 - Mtry = 80
 - Node size = 1
 - k = 75 → (Number of clusters)
- Total: RFSRC w/ k-medoid
 - Mtry = 22
 - Node size = 10
 - k = 75