Using Machine Learning to Understand Predictors of Frequent Surveillance Testing in Colorectal Cancer Patients

Ronald Dickerson, MA, MPH[1]; Eric Haupt, ScM[2]; Eric Hahn, MPH, Ph.D[2].; Aasthaa Bansal, Ph.D.[1]
[1]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington
[2]Kaiser Permanente Southern California

UNIVERSITY of WASHINGTON
THE CHOICE INSTITUTE
School of Pharmacy

Background

- Post-treatment surveillance is crucial for colorectal cancer (CRC) patients to detect recurrence early.
- Guidelines recommend regular carcinoembryonic antigen (CEA) testing every three to six months for five years post-treatment.
- However, patients often do not receive recommended surveillance testing. Understanding real-world surveillance patterns may help enhance post-treatment care for CRC survivors
- At a given surveillance visit, if providers could identify patients who have a low likelihood of returning for the next visit on time, they could potentially intervene. This presents an opportunity to improve adherence to guidelines and enhance patient care.

Objectives

- This study aims to identify key patient and physician characteristics that predict the likelihood of a patient returning for frequent CEA testing in current practice.
- By accurately predicting the probability of a patient returning for the next visit on time, healthcare providers can intervene on patients who have a low likelihood of returning for their next surveillance visit.

Methods

Data Source: De-identified Kaiser Permanente Southern California EHR database.

Patients: 2,489 adult patients diagnosed with AJCC stage I-III CRC between 2008 and 2013.

Outcome: At each visit, we defined the binary outcome of whether the next surveillance CEA test occurred within three months. We treated multiple testing visits during surveillance as a recurrent event.

Candidate Features and Feature Engineering:

- Baseline and time-varying patient variables, such as prior and current CEA values
- Created time-varying covariates for time since last surveillance CEA visit, time since surveillance start, and time since last surveillance imaging
- Derived biomarker-related summary measures, such as trend in biomarker levels over time

Model Fitting: Utilized Logistic Regression with Lasso Penalty. Conducted variable selection among 252 patient and physician characteristics.

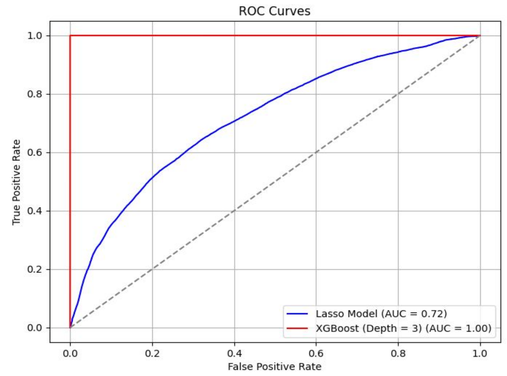
Internal Validation: Split-sample validation approach, dividing data into 75% training and 25% validation sets.

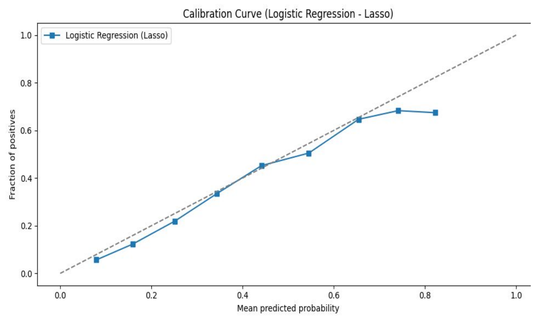
Cross-validation: Determined optimal regularization parameters using 10-fold cross-validation within the training set.

Model Evaluation: ROC curves and calibration plots.

Interpretability: Conducted feature importance analyses to enhance model interpretability and identify key predictors.

Results

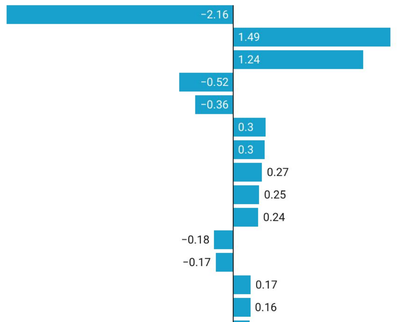




- The Lasso model has good discriminatory power in predicting patients' likelihood of undergoing CEA testing within the next three months. Importantly, the **Lasso** model demonstrates good calibration, indicating close alignment between predicted probabilities and observed outcomes.
- Patients were less likely to undergo testing within three months if it had been longer since the start of surveillance and if they were diagnosed with Stage 1 or 2 disease. They were more likely to undergo testing if it had been longer since last imaging and they had a high prior CEA value.

Lasso

Penalized Regression: Variable Selection (Average Beta Coefficients)



Conclusions

- This research sheds light on current surveillance patterns and may be helpful in designing interventions that deliver appropriate, value-based post-treatment care for CRC survivors
- By accurately predicting the probability of a patient returning for the next visit on time, this model could also be used in practice by providers to intervene on patients who have a low likelihood of returning for their next surveillance visit.

Acknowledgements

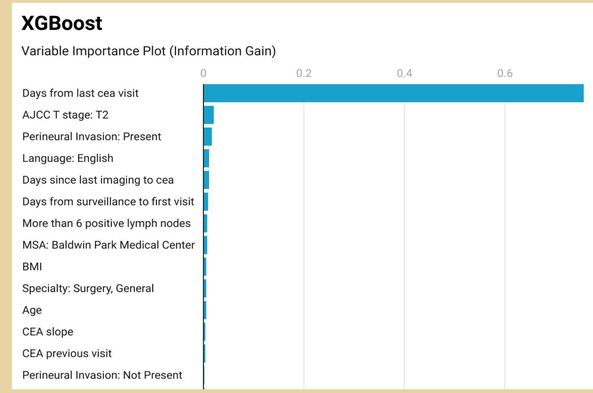
- This study was funded by the National Cancer Institute of the National Institutes of Health (NIH) under grant R37-CA218413: Personalized Risk-Adaptive Surveillance in Cancer (PRAISE).

- **Timing between between surveillance and CEA testing, CEA levels, and stage**

QUICK FACTS

Study Title: Using Machine Learning to Understand Predictors of Frequent Surveillance Testing in Colorectal Cancer Patients

- Study Type: Retrospective Cohort Study
- Data Source: Kaiser Permanente Southern California EHR Database
- Patient Population: Adult patients diagnosed with AJCC stage I-III CRC between 2008 and 2013
- Outcome of Interest: Biomarker testing during surveillance, treated as a recurrent event
- Analytical Methods: Logistic Regression with Lasso Penalty and XGBoost models
- Key Findings: Lasso model moderately predicts likelihood of frequent CEA testing; XGBoost model demonstrates strong accuracy
- Implications: Insights can guide post-treatment care strategies for CRC survivors



- **XGBoost** demonstrates strong accuracy in identifying patients who are likely to undergo CEA testing within the next three months, with an exceptional AUC. The calibration plot reveals a deviation from the expected calibration, suggesting potential challenges with the model's calibration. Future work will consider recalibration techniques to improve the alignment between