Cross-validation of Machine Learning for Five Year All-Cause Mortality Prediction in Prostate and Breast Cancer

Huan-ju Shih, MS; Janusz Wojtusiak, PhD; Hua Min, PhD
Health Informatics Program, College of Public Health, George Mason University

1. Introduction

- Advancements in detection increase survival rates for prostate & breast cancer → focus on HRQOL & survivorship predictors.
- Few studies compare 1 year & 5 years cancer-specific/all-cause mortality prediction models across cancer types using integrated SEER-MHOS data.
- Evaluate within-cancer models + cross-testing (prostate & breast) for generalizability.

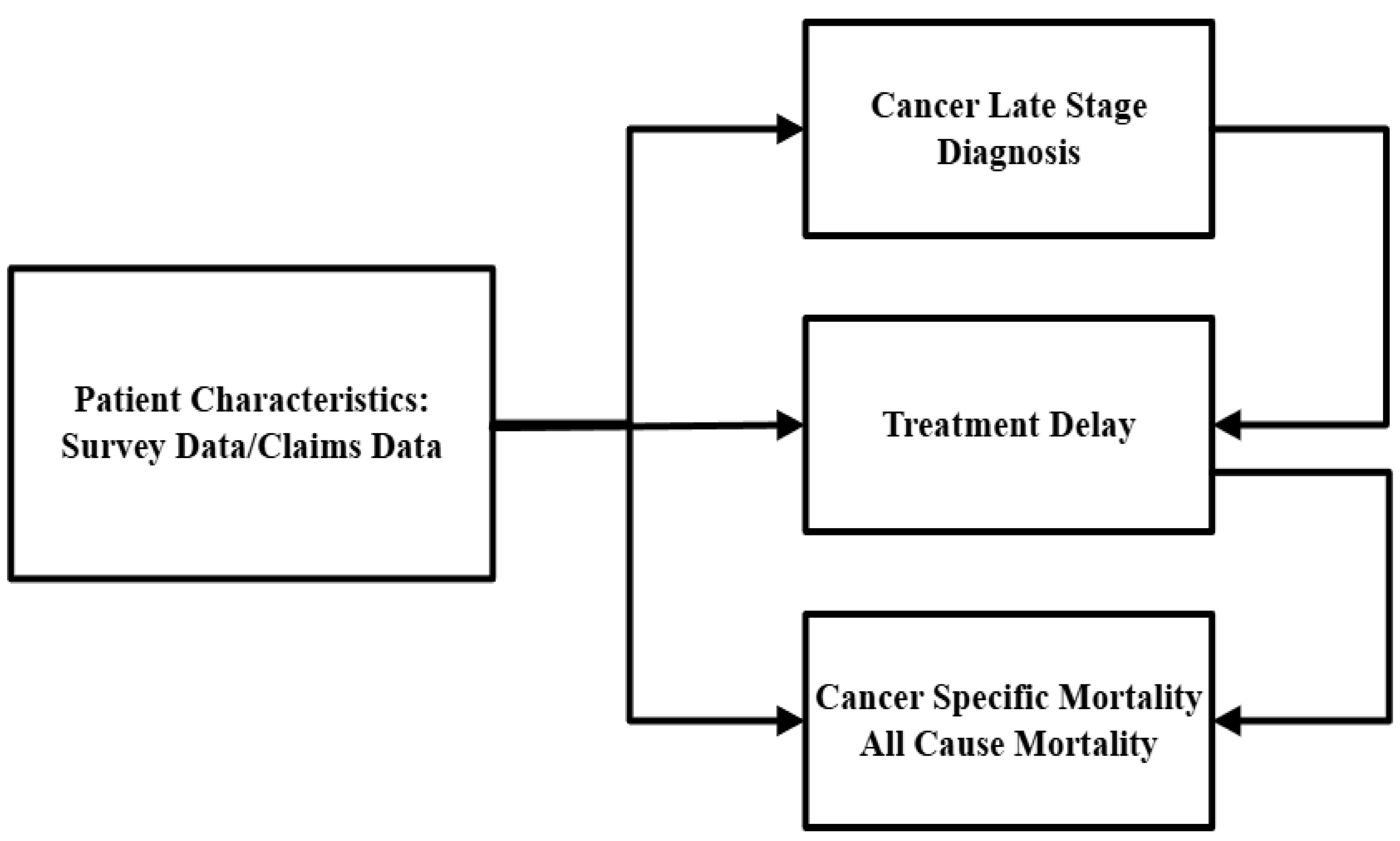
2. Data Collection - Methods

- Integrates NCI SEER registry + CMS Medicare Health Outcomes Survey.
- Prostate & breast cancer patients ≥65 yrs (1998–2011).

Experimental Framework

- Process I: Imputation plus optimal lookback window selection.
- Process II: Model training, subgroup analysis (race, gender, ethnicity), cross-testing.

3. Data Integration - Workflow & Results



Uses SEER-MHOS survey data to link patient factors (demographics, comorbidities, function, access) to late-stage diagnosis, treatment delays, and mortality.

Model: Column; Test Set: Row

Outcomes	One Year All Cause Mortality			Five Years All Cause Mortality		
Cancer Type	Prostate	Breast	Combine	Prostate	Breast	Combine
Prostate	0.78	0.74	0.78	0.80	0.75	0.79
Breast	0.74	0.73	0.75	0.79	0.79	0.79
	One Year Cancer Specific Mortality			Five Years Cancer Specific Mortality		
Cancer Type	Prostate	Breast	Combine	Prostate	Breast	Combine
Prostate	0.74	0.74	0.85	0.74	0.73	0.76
Breast	0.74	0.80	0.81	0.70	0.73	0.75

Random Forest models achieved the highest AUCs (up to 0.85) for predicting mortality in prostate cancer patients, with strong performance across 1-year and 5-year all-cause and cancer-specific outcomes.

4. Data Processing & Analysis - Workflow & Results

Cross-Testing I

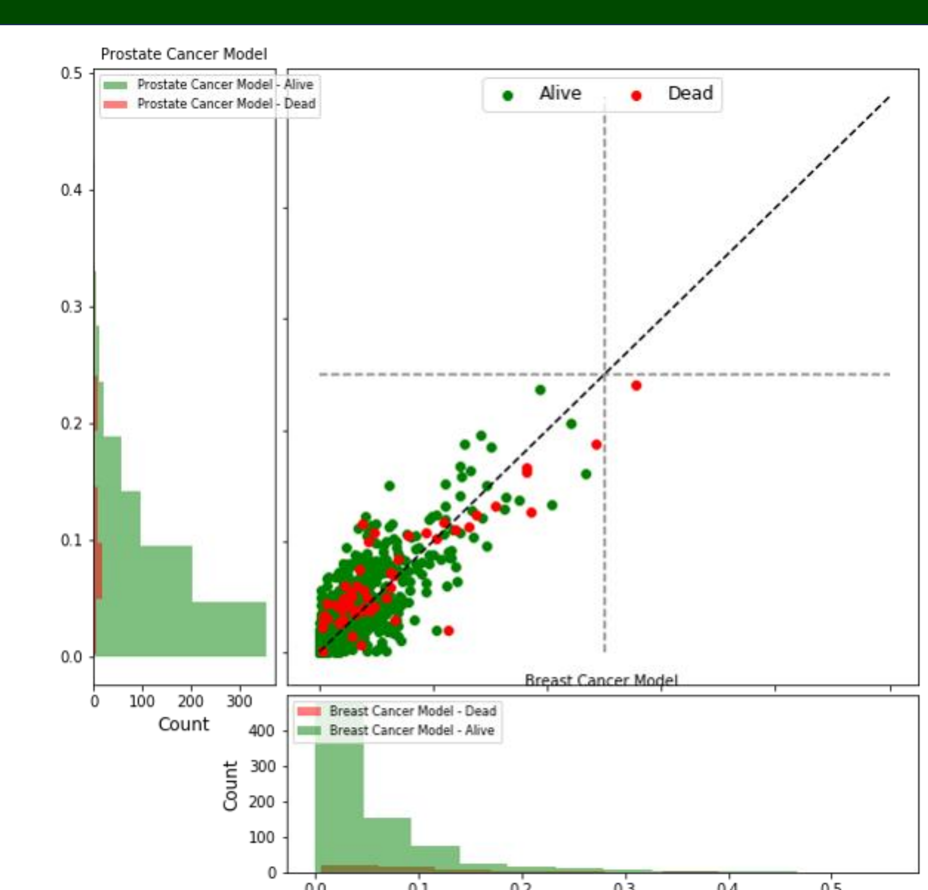


Figure 1. One Year All Caused Mortality. Comparison of Prostate and Breast Cancer Models Applied Breast Test Set (plot 1)

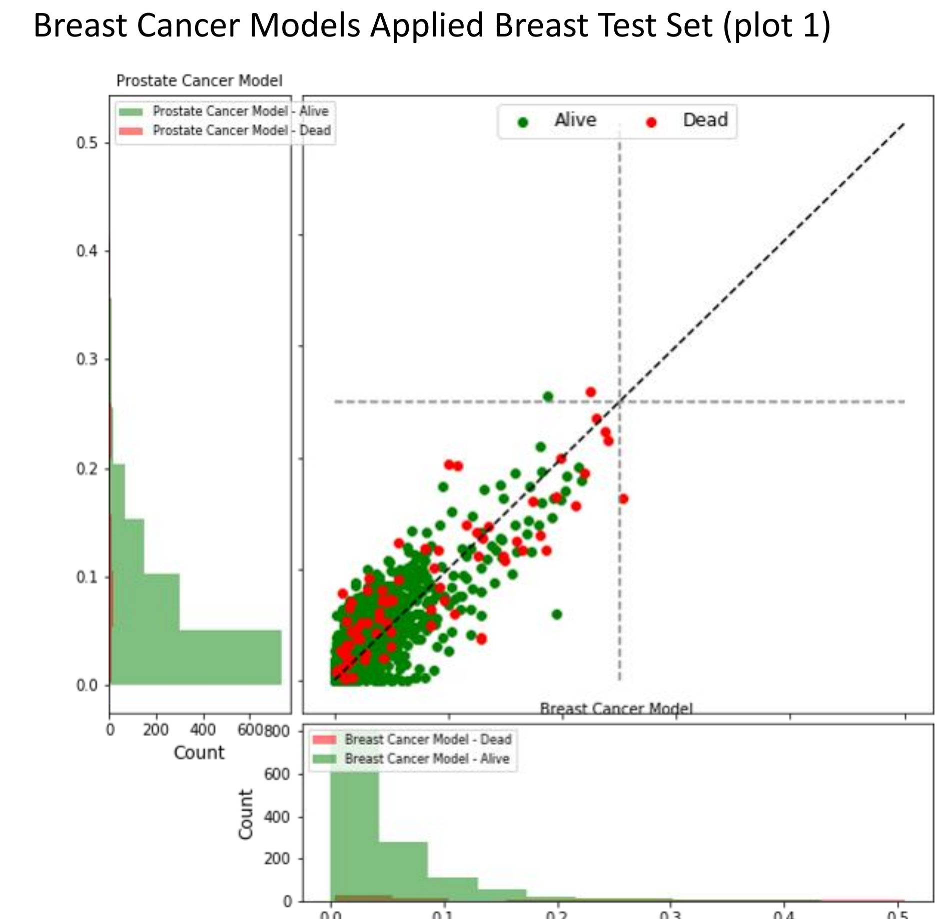


Figure 2. One Year All Caused Mortality. Comparison of Breast and Concatenate Cancer Models Applied Breast Test Set (plot 2)

Cross-Testing II

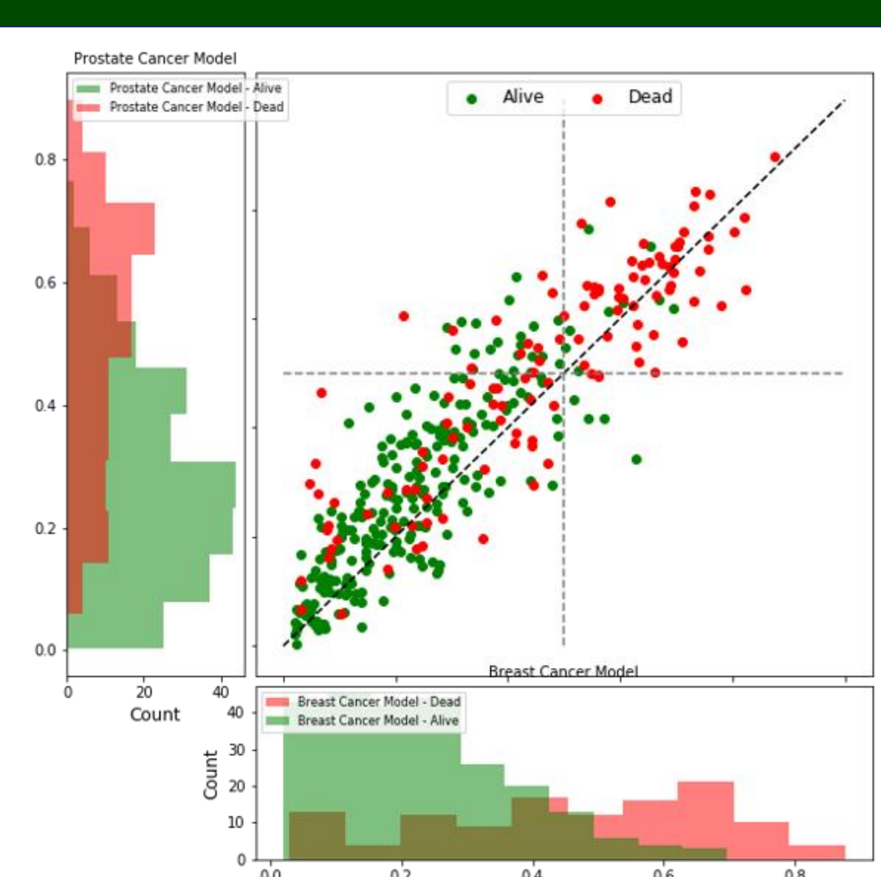


Figure 3. Five Years All Caused Mortality. Comparison of Breast and Prostate Cancer Models Applied Breast Test Set (plot 1)

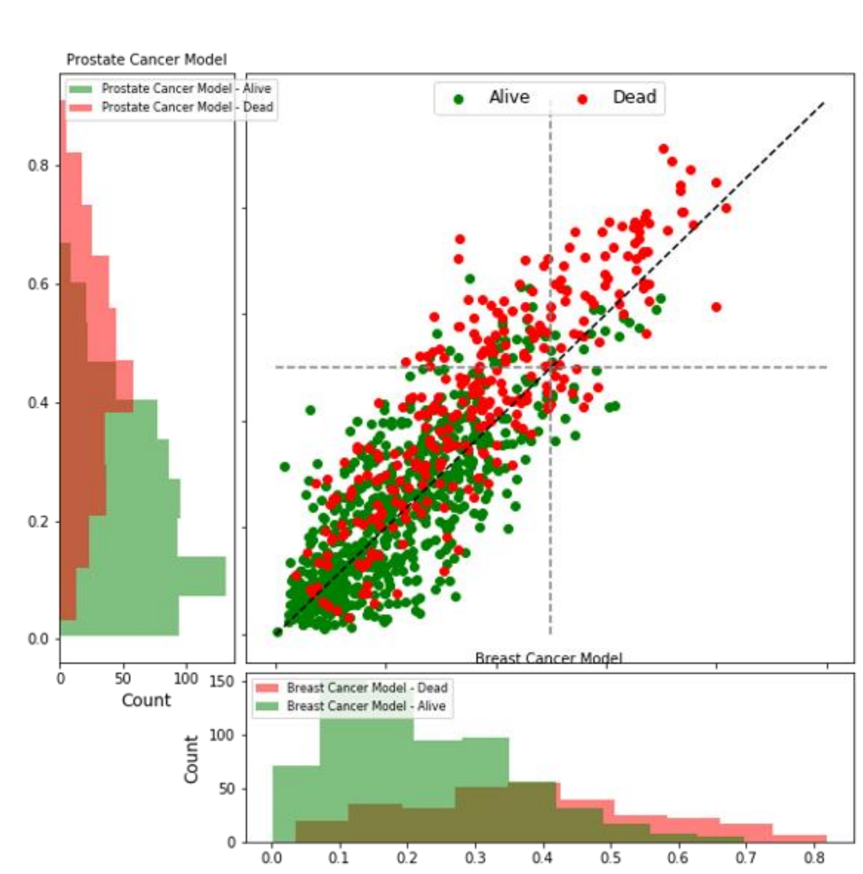
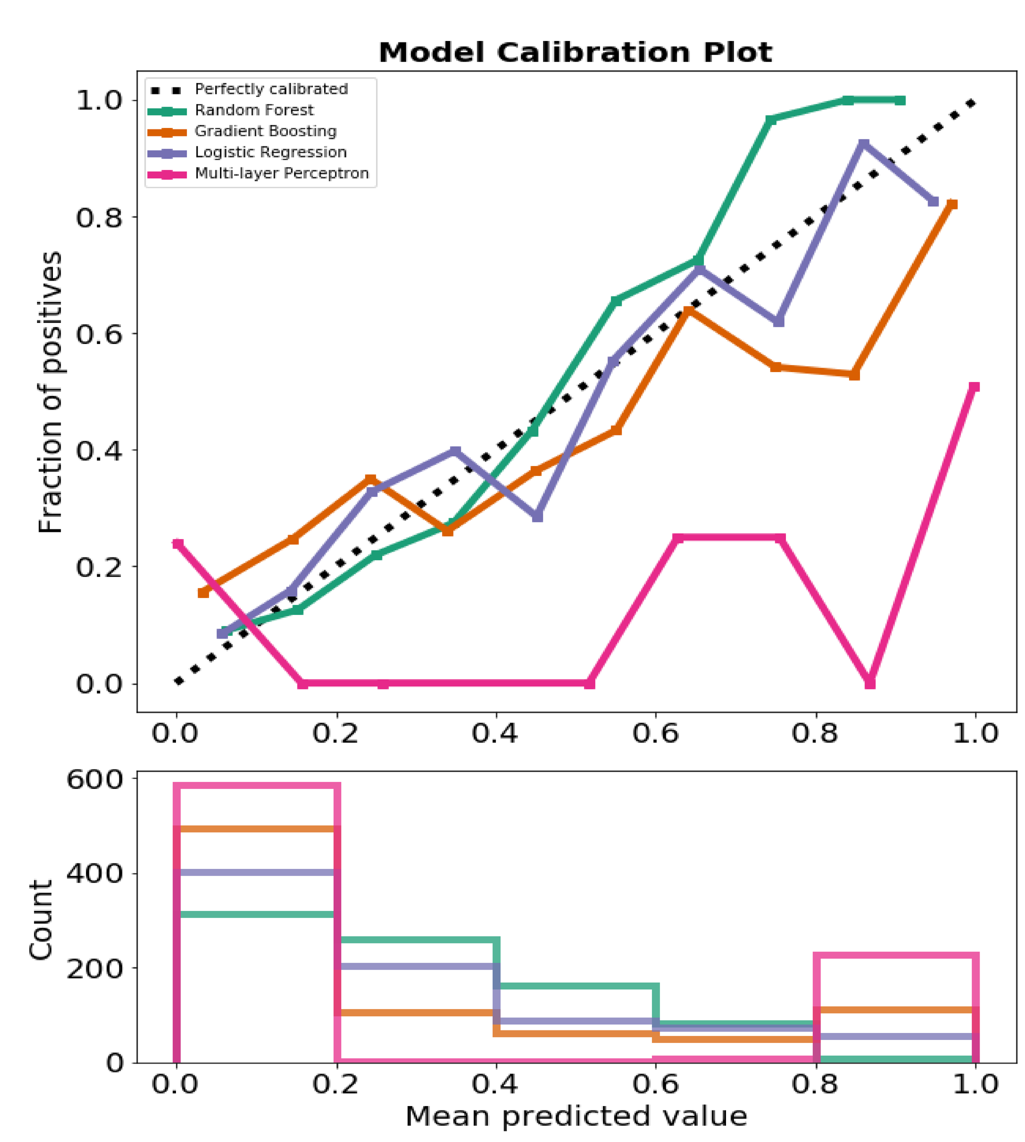


Figure 4. Five Years All Caused Mortality. Comparison of Breast and Prostate Cancer Models Applied Prostate Test Set (plot 2)

Calibration

Figure 5. SEER-MHOS Calibration Concatenate Cancer Five Years All Cause Mortality



5. Conclusion

- Combined data increase stability
- /accuracy; reveals shared predictors but risks bias in high-risk groups.
- Cross-testing informs context-specific oncology modeling; supports data-driven survivorship care.