# **Long-Term Survival Prediction in Early Breast Cancer** : A Machine Learning Approach with Random Survival Forest

Hyuna Yoon<sup>1,2,3</sup>, Sola Han<sup>1</sup>, Hae Sun Suh<sup>2,3,4</sup>, Chanhyun Park<sup>1</sup> 1. College of Pharmacy, The University of Texas at Austin; 2. College of Pharmacy, Kyung Hee University; 3. Institute of Regulatory Innovation through Science, Kyung Hee MSR49 University, Seoul, Republic of Korea; 4. Department of Regulatory Science, Graduate School, Kyung Hee University.

## BACKGROUND

- Breast cancer (BC) is the most common cancer among women, and it has the second-highest mortality rate among all cancers in women in the US.<sup>1</sup>
- Regarding the causes of death among patients with breast cancer, the primary cause is breast cancer itself, and among non-cancer related causes of death, cardiovascular disease is the most common.<sup>2,3</sup>
- Predicting breast cancer survival curves can help improve patient outcomes.
- The Random Survival Forest (RSF) algorithm showed good predictive performance in predicting breast cancer patient's death.4,5

## **OBJECTIVES**

- To develop survival prediction models for all-cause, BC-caused, and cardiovascular disease (CVD)-caused mortality in older women with hormone receptor-positive (HR+) early BC.
- To identify key prognostic factors for mortality.

## METHODS



- **Data Source:** SEER-Medicare
- **Study Population**
- Female
- Age 66 years old and older
- Early-stage breast cancer Hormone receptor positive
- Had primary surgery for breast cancer including breast conserving surgery

**Index Date 1st use of Adjuvant Endocrine Therapy (AET)** Day 0

• Part A/B/D: continuous enrollment needed HMO: excluded if enrolled at any point

Pre-index: 12 month before index date

**Diagnosis of breast** cancer (C50)

Post-index: 10-year following the index date or until an outcome event



(2006.01.01)

(2007.01.01~2009.12.31)

occurs

(2019.12.31)

Model Development: Random Survival Forest

- Training set (75%) / Test set (25%)
- Variables: Demographic (age, race), Cancer-related (tumor size, SEER summary stage), Diagnosis-based comorbidity defined using Clinical Classification Software (CCS) (285 categories)
- Feature Selection: LASSO method in Cox Model
- Hyper-parameter tuning
- Number of estimators: (10, 100, 500, 1000)
- Minimum of samples split: (2, 3, 5, 10, 15, 20)
- Minimum of samples leaf: (1, 3, 4, 10, 15, 20)

	RESULTS									
1	Figure 1. Model Perform	Figure 1. Model Performance								
		Old (66-79 years old) Patient				Oldest old (≥ 80 years old) Patient				
		All cause death	BC-related death			All cause death	BC-related death	CVD related death		
	C-index	0.719	0.789	0.742		0.693	0.756	0.707		
	Mean AUROC	0.771	0.811	0.737		0.763	0.805	0.765		
	IBS	0.109	0.047	0.016		0.154	0.073	0.054		
	Figure 2. SHAP Plots for A				-					
		Index date age			Index date age					
	Other screening for suspected conditions(not	t mental disorders or infectious disease)			Other screening for suspected conditions(not mental disorders or infectious disease)					
	Cong	Congestive heart failure; non-hypertensive			Congestive heart failure; non-hypertensive Delirium, dementia, and amnestic and other cognitive disorders					
	Chronic obstructive	Cancer stage (Localized / Regional)				Cancer stage (Localized / Regional)				
	Chronic obstructive pulmonary disease and bronchiectasis Secondary malignancies				Size $\geq 2.0$ cm					
ų.	Size ≥ 2.0 cm				Chronic obstructive pulmonary disease and bronchiectasis					
	Delirium, dementia, and amnestic and other cognitive disorder				Immunizations and screening for infectious disease Disorder of lipid metabolism					
L	Respiratory failure; insufficiency; arrest (adult)			<u>e</u>	Radiation therapy					
	Deficiency and other anemia			e Ka						
	66-79 age group				80+ age group					
	Figure 3. SHAP Plots for BC-Cause Death							High		
	Size $\geq 2.0$ cm				Cancer stage (Localized / Regional)					
		Cancer stage (Localized / Regional)			Other screening for suspected conditions(not mental disorders or infectious disease) Size ≥ 2.0 cm					
		Secondary malignancies			Index date age					
		Progesterone hormone receptor			Secondary malignancies					
	Other screening for suspected conditions(not mental disorders or infectious disease)			1000000	Immunizations and screening for infectious disease					
	Index date age				Congestive heart failure; non-hypertensive Deficiency and other anemia					
	Other and unspecified benign neoplasm				Cataract					
		Fluid and electrolyte disorders			Progesterone hormone receptor					
	Other inju	Other injuries and conditions due to external causes			<u>e</u>					
		Deficiency and other anemia			80+ age group					
	Eiguro A SHAD Plate for C	66-79 age group			Index date age					
	Figure 4. SHAF Flots for C	Figure 4. SHAP Plots for CVD-Cause Death			Congestive heart failure; non-hypertensive					
		Congestive heart failure; non-hypertensive			Cataract					
		Index date age				Acute cerebrovascular disease Other screening for suspected conditions(not mental disorders or infectious disease)				
	Peri-; endo-; and mvocarditis: cardiomvopat	Chronic obstructive pulmonary disease and bronchiectasis				Start Servening for Suspected conditions(not incluta	Heart valve disorders	value		
	,,,,,,,,,	Other and ill-defined heart disease				Oth	ner and ill-defined heart disease	eature		
	Pulmonary heart disease					Oth	ner upper respiratory infections			

- Diabetes mellitus with complications
- Coronary atherosclerosis and other heart disease
- Other injuries and conditions due to external causes
  - Benign neoplasm of uterus
  - Cardiac dysrhythmias

## 66-79 age group

## Strengths

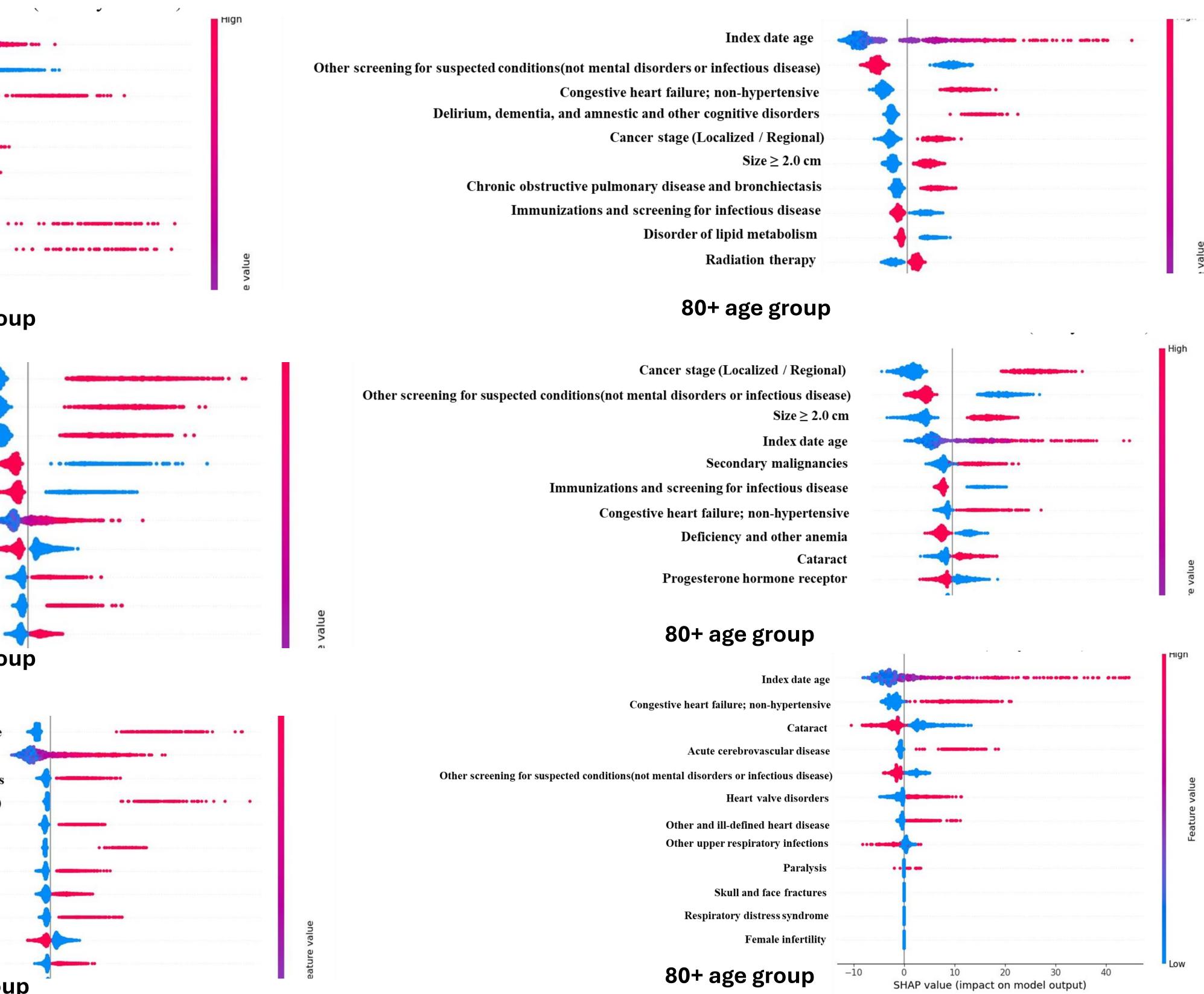
- This study included patient's comorbidities, which are not included in the prior studies, as the model predictors.
- This study explained the developed model using SHAP and identified the factors that influence patients' death.

## Conclusion

- algorithm.
- Age, screenings, and congestive heart failure were key predictors of mortality of all cause death regardless of age.







## CONCLUSION

This is the study that developed 10-year long-term survival prediction models of the older patients with breast cancer for the U.S. popu

• We developed survival prediction models for older women with breast cancer which showed acceptable performances using random

Cancer characteristics such as tumor size, cancer stage, and presence of secondary malignancies were the key factors for BC-caused Congestive heart failure, heart valve disorder, and other ill-defined heart diseases were the key predictors for the CVD-caused death.

# The University of Texas at Austin **Division of Health Outcomes**



## REFERENCE

ulation.	<ol> <li>Siegel, R. L. et al., CA Cancer J Clin. 2023; 73(1), 17-48.</li> <li>Afifi, A.M. et al., Cancer. 2020; 126(7), 1550, 1567</li> </ol>				
forest d mortality.	<ul> <li>126(7), 1559-1567.</li> <li>3. Reddy, P. et al., CA Cancer J Clin.</li> <li>2017;8(4) 308-314 .</li> <li>4. Najafi-Vosough, R. et al., J Prev</li> <li>Med Hyg. 2022; 63(2).</li> <li>5. Rahman, M. S. et al., ICECET.</li> <li>2023.</li> </ul>				
	Contact: hyuna.yoon@austin.utexas.edu				