Breast cancer is a leading cause of death among women. We develop a continuous risk model incorporating easily accessible EHR information.

Developing an Interpretable Breast Cancer Prediction Model Incorporating Demographic, Drug, and Disease Information

Matthew Murrow, PhD, Congning Ni, PhD, Sharla Rahman, Xingyi Guo, PhD, Zhijun Yin, PhD

Introduction

Globally effects millions of women. >2 million new diagnoses every year [1].

Leading cause of cancer related mortality. Over half a million women die every year [2].

Easiest to treat if caught early! Early access to diagnoses reduces mortality by 25-40% [3].

US recommends bi-yearly mammograms for women aged 40-74 at average risk [4].

Women known to be at high risk (e.g. family history or genetic factors) recommended to receive more frequent screening [4].

Why build a risk model?

Mammograms are imperfect and subjective. False negatives do happen!

Inaccurate on women with dense breast tissue.

Not everyone has access to bi-yearly mammograms!
- Live far from a hospital.
- Lack health insurance or underinsured.
- May not have transportation.
- Don't think you need it (e.g. negative first mammogram, no obvious risk factors, first mammogram negative).

Treatment recommendations tailored to populations, not individuals.

Routine screening and cancer treatment are expensive!

In this work

- Develop a breast cancer risk model incorporating demographic, drug, and disease diagnoses.
- Targeted towards women who receive negative mammograms and are not flagged for enhanced screening.
- Utilize a landmark modeling approach. Estimate breast cancer risk starting from first mammogram.

Cohort

Retrospective cohort built from Vanderbilt University Medical Center's EHR synthetic derivative.

Includes women who:
1. Received ≥2 mammograms (on different days).
2. First and final mammogram ≥1 year apart.
3. First mammogram must be negative (BI-RADS 1 or 2).
4. No previous cancer diagnosis (of any type).

Cohort contains 52,929 women. Of those, 2,031 eventually received a breast cancer diagnosis.

Features

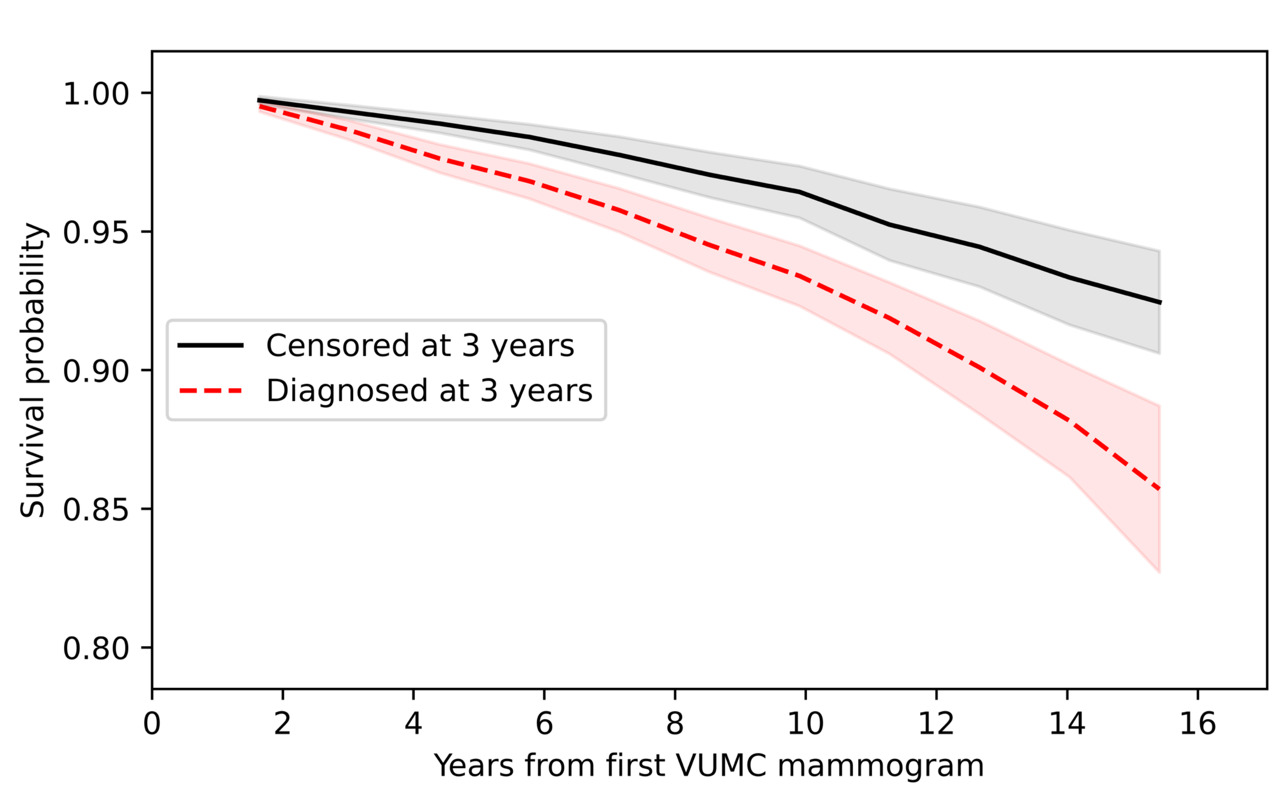
Demographics including age, BMI, menopause status, family breast/ovarian cancer history, race, and ethnicity.

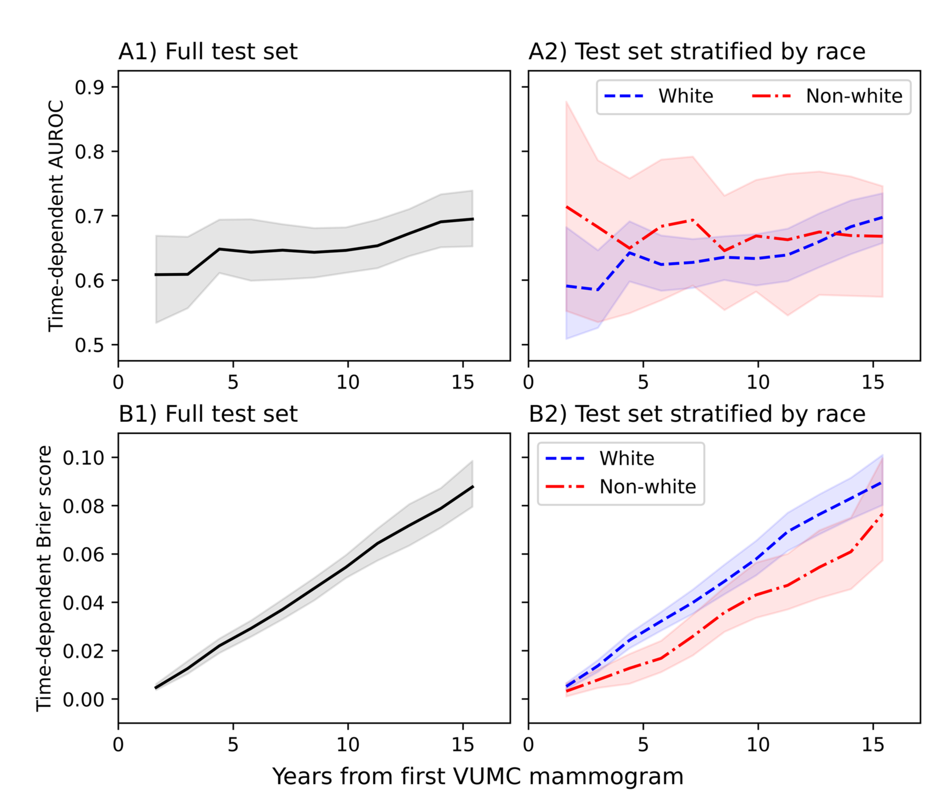
Drug exposures prior to first mammogram. Mapped to Anatomical Therapeutic Chemical (ATC) code classification.

Disease diagnoses prior to first mammogram. Identified using ICD codes.

Model

Landmark model predicting survival probability from first mammogram using Survival Random Forest.





VANDERBILT UNIVERSITY
MEDICAL CENTER

Department of
Biomedical Informatics

Take a picture to download poster

Supported by NLM T15 Training Grant

matthew.murrow@vumc.org