A Machine Learning Approach for Identifying Breast Cancer Recurrence Events in Population-based Claims Data

Lucas J. Liu[1], Quan Chen[2], A'mar, Teresa[1], Ruth B. Etzioni[1], Val R. Adams[2], Jessica Chubak[3], Qi Qiao[2], Jin Chen[2], Bin Huang[2]

[1]Fred Hutchinson Cancer Center, Seattle, WA, [2]University of Kentucky, Lexington, KY, [3]Kaiser Permanente Washington Health Research Institute, Seattle, WA

Fred Hutch Cancer Center

Introduction

- Cancer recurrence is an important outcome for measuring illness burden and treatment efficacy, however cancer recurrence is not explicitly documented in US cancer registries[1].
- Studies show promising results in inferring cancer recurrence from claims data, but primarily on small or single-institution studies.
- We developed a supervised machine learning framework to identify the occurrence and the timing of the second breast cancer events (SBCE) using registry and medical claims data at population level.

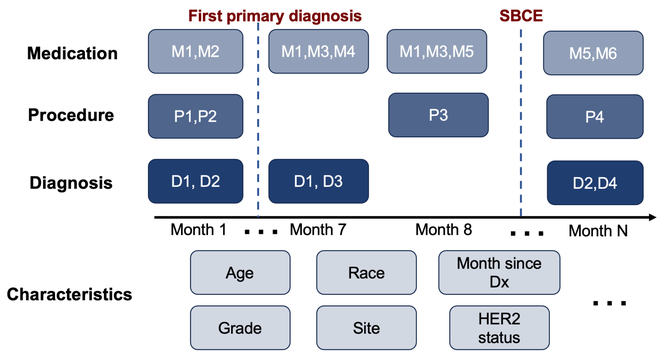
Data

- This study includes **18239** adult female patients diagnosed with first primary, stages I,II,III, local or regional breast cancer between 2004 and 2015 in Kentucky.
- Data are collected from Kentucky Cancer Registry (KCR) linked with Medicaid and Medicare claims. The SBCE rates were **7.25%** in the study cohort.
- **SBCE definition**:
  1) Recurrence of the first primary breast cancer **or**
  2) Diagnosis of the second primary breast cancer **or**
  3) Death caused by the first primary breast cancer
- **Features**:
  - **22** patients' characteristics (e.g., age, etc.) and first primary breast cancer characteristics (e.g., tumor size, etc.)
  - **29811** ICD9/10 diagnosis codes
  - **17718** ICD9/10 and HCPCS procedures codes
  - **41271** NDC and AHFS drug codes

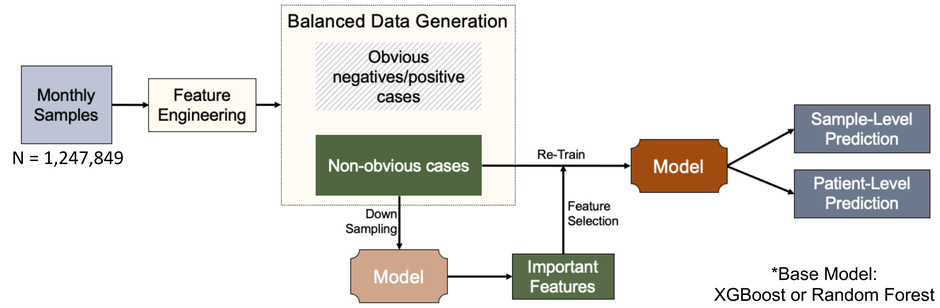
Abbreviation :
ICD: The International Classification of Disease
HCPCS: Healthcare Common Procedure Coding System
NDC: The National Drug Code
AHFS: American Hospital Formulary Service Pharmacologic-Therapeutic Classification System

Methods

Patient Data



Overall Framework



Results

Table 1. Patient-level prediction performance on test data

Threshold	Accuracy	Sensitivity	Specificity	*Difference
0.1	0.85	0.77	0.86	-6.0 ± 17.1
0.2	0.87	0.72	0.88	-5.1 ± 15.5
0.3	0.88	0.69	0.90	-5.0 ± 15.7
0.4	0.89	0.68	0.91	-4.7 ± 15.4
0.5	0.90	0.66	0.92	-3.8 ± 17.3
0.6	0.90	0.63	0.93	-3.1 ± 17.7
0.7	0.91	0.61	0.93	-2.5 ± 18.5
0.8	0.91	0.56	0.94	-1.4 ± 19.3
0.9	0.92	0.50	0.96	0.9 ± 19.2

*Raw difference (mean ± SD) between the predicted and observed time of SBCE
** ROCAUC is 0.93 at sample-level prediction

For more information, please contact Lucas J. Liu at jliu6@fredhutch.org

Results, cont.

Figure 1. Example of Prediction Trajectory

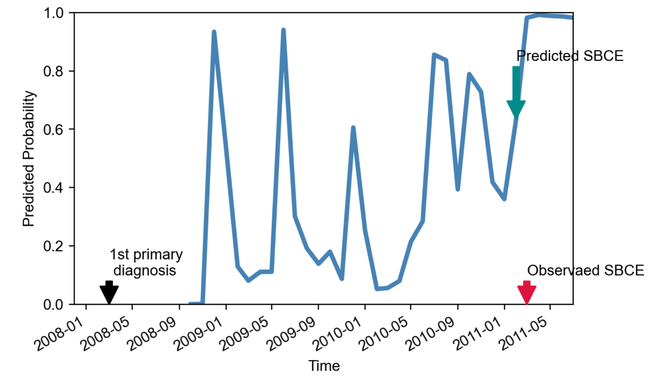
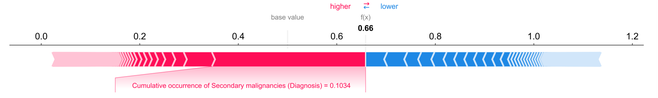


Figure 2. Individual SHAP[2] Interpretation



- Cumulative occurrence of diagnosis secondary malignancies contributes the most to the prediction

Conclusions

- Our machine learning framework identified SBCE using claims data with high ROCAUC, accuracy, and specificity at the population level.
- It is important to validate this framework in other populations' datasets for further confirmation of its effectiveness.

Acknowledgements

- NCI UH3CA218909 ReCAPSE: Recurrence from Claims And PROs for SEER Enhancement.

References

1. Warren, Joan L., and K. Robin Yabroff. "Challenges and opportunities in measuring cancer recurrence in the United States." Journal of the National Cancer Institute 107.8 (2015): djv134.
2. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. Advances in neural information processing systems 30, 2017