Machine Learning-Based Explanation of Racial–Ethnic Disparities in 5-Year Cancer-Specific Survival Among Hormone Receptor-Positive Breast Cancer Patients in the United States

CO65

Rashed Harun, Eunice Kim, Anna Cheng, Ibrahim M Abbass

Genentech, Inc., South San Francisco, CA, USA

BACKGROUND

- Several disparities in breast cancer outcomes between patients with different racial backgrounds have been reported:
  - Black women continue to have higher mortality than White women.[1–4]
  - Black women are also diagnosed at later cancer stages and have larger tumor sizes at diagnosis.[1]
  - Black women are less likely to obtain adequate treatment and care, including experiencing more delays, than White women.[2, 3, 5, 6]
- Survival outcomes in patients with breast cancer can be predicted by machine learning (ML).
  - ML has been shown to be similar to, and in some cases better than, Cox proportional hazards analysis at predicting survival.[7]
  - However, previous lack of explainability (understanding the reasoning behind the ML predictions) hindered wider adoption in clinical practice.[7,8]
- We developed an ML model to predict 5-year cancer-specific survival of patients with hormone receptor-positive (HR+) breast cancer and to quantify the drivers of racial–ethnic disparities.

METHODS

- This was a retrospective cohort study using Surveillance, Epidemiology, and End Results (SEER) Program cancer registry data and county-level Socioeconomic Status (SES) Index data from the Centers for Disease Control and Prevention.
- The study population, inclusion criteria, and outcomes are shown in **Table 1**.
- The 5-year cancer-specific survival was predicted using an XGBoost ML model using 12 variables related to cancer characteristics, surgery, and demographic factors.
- The model's performance was measured using concordance index (C-index) using 5-fold cross-validation.
- SHapley Additive exPlanations (SHAP) analysis was used to explain the impact of specific variables on hazard predictions, including across racial–ethnic subgroups.

Table 1: Study population, inclusion criteria, and outcomes

Population	Patients ≥18 years at the time of HR+ breast cancer diagnosis in the SEER cancer registry	
		n (%)
Inclusion criteria and attrition numbers	• No histologic codes of 9050–9055, 9140, or 9590–9992	551,222 (100)
	• Receptor subtype = HR+, HER2– (Luminal A) or HR+, HER2+ (Luminal B)	349,549 (63.4)
	• Malignant cases only	340,627 (61.8)
	• Cancer is microscopically confirmed	340,556 (61.8)
	• Cancer is first primary: "one primary only" or "first of two or more primaries"	274,732 (49.8)
	• Active follow-up with and reporting source NOT "autopsy only" and NOT "death certificate only"	274,732 (49.8)
	• Excludes cases with unknown cancer subtype, grade, or stage	259,305 (47.0)
Study period	January 2010–December 2016	
Follow-up period	From the date of diagnosis date until the earlier occurrence of death or study end	
Outcomes	5-year cancer-specific survival	

HER2–, HER2-negative; HER2+, HER2-positive; HR+, hormone receptor-positive; SEER, Surveillance, Epidemiology, and End Results.

RESULTS

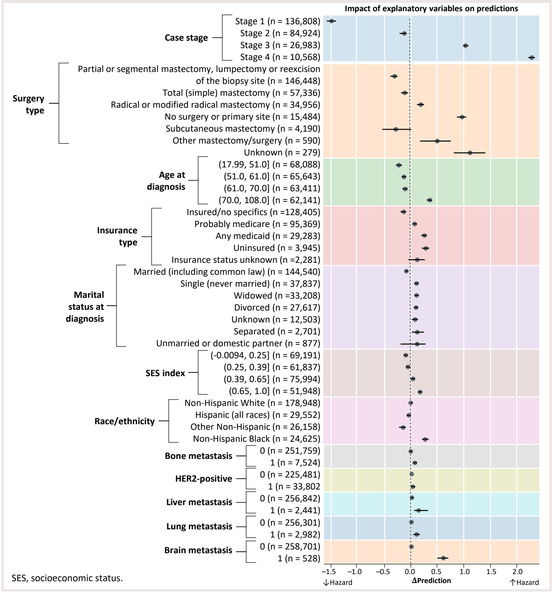
- Following application of the selection criteria, the study population included 259,305 patients with HR+ breast cancer (**Table 1**).
- Of these, 69.0% were Non-Hispanic White, 11.4% were Hispanic, 9.5% were non-Hispanic Black, and 10.1% were Other Non-Hispanic. Baseline characteristics by race/ethnicity are shown in **Table 2**.
- The ML model predicted 5-year cancer-specific survival well, with a C-index of 0.89±0.003.
- Of what could be explained of cancer-specific survival hazards, cancer staging explained 50.7% of the variation, followed by surgery type (15.1%), age at diagnosis (10.7%), insurance type (6.5%), marital status (4.9%), SES (4.5%), and race/ethnicity (3.2%) (**Figure 1**).
- Compared with Non-Hispanic White patients, mortality risk was similar in Hispanic patients; however, it was higher in Non-Hispanic Black patients (hazard ratio 1.55 [95% confidence interval 1.42, 1.69]) (**Figure 2**).
- The latter racial–ethnic disparity was largely mediated by differences in cancer stage at diagnosis (36.7%), surgery type (10.8%), marital status (6.5%), SES (6.2%), and insurance type (4.9%) (**Figure 3**).

Table 2: Baseline characteristics

Characteristic	Hispanic (all races) n = 29,554	Non-Hispanic Black n = 24,625	Non-Hispanic White n = 178,963	Other Non-Hispanic n = 26,163
Age at diagnosis	56 (47, 66)	59 (49, 68)	62 (53, 71)	57 (48, 67)
Marital status at diagnosis				
Married*	16,027 (54.2)	8,318 (33.8)	103,791 (58.0)	16,411 (62.7)
Single†	11,935 (40.3)	14,902 (60.5)	66,308 (37.1)	8,221 (31.4)
Unmarried or domestic partner	104 (0.4)	50 (0.2)	656 (0.4)	67 (0.3)
Unknown	1,488 (5.0)	1,355 (5.5)	8,208 (4.6)	1,464 (5.6)
Region				
West	23,980 (81.1)	6,421 (26.1)	87,593 (48.9)	22,343 (85.4)
Midwest	438 (1.5)	2,767 (11.2)	19,039 (10.6)	533 (2.0)
Northeast	3,745 (12.7)	3,783 (15.4)	30,844 (17.2)	2,278 (8.7)
South	1,391 (4.7)	11,654 (47.3)	41,487 (23.2)	1,009 (3.9)
Insurance status				
Insured‡	28,205 (95.4)	23,661 (96.1)	175,783 (98.2)	25,424 (97.2)
Uninsured	1,051 (3.6)	685 (2.8)	1,771 (1.0)	438 (1.7)
Unknown	298 (1.0)	279 (1.1)	1,409 (0.8)	301 (1.2)
Breast cancer receptor subtype				
HR+, HER2– (Luminal A)	24,994 (84.6)	20,640 (83.8)	157,584 (88.1)	22,284 (85.2)
HR+, HER2+ (Luminal B)	4,560 (15.4)	3,985 (16.2)	21,379 (11.9)	3,879 (14.8)
Cancer stage				
Stage 1	13,504 (45.7)	10,831 (44.0)	99,260 (55.5)	13,226 (50.6)
Stage 2	10,895 (36.9)	8,931 (36.3)	55,785 (31.2)	9,318 (35.6)
Stage 3	3,963 (13.4)	3,303 (13.4)	17,009 (9.5)	2,711 (10.4)
Stage 4	1,192 (4.0)	1,560 (6.3)	6,909 (3.9)	908 (3.5)
Cancer grade				
Grade I or II	21,150 (71.6)	16,298 (66.2)	139,734 (78.1)	19,117 (73.1)
Grade III or IV	8,404 (28.4)	8,327 (33.8)	39,229 (21.9)	7,046 (26.9)
Surgery type				
No surgery of primary site	2,036 (6.9)	2,256 (9.2)	9,667 (5.4)	1,532 (5.9)
Other mastectomy/surgery	106 (0.4)	51 (0.2)	372 (0.2)	61 (0.2)
Partial or segmental mastectomy, lumpectomy, or re-excision of the biopsy site	15,272 (51.7)	12,893 (52.4)	104,971 (58.7)	13,322 (50.9)
Radical or modified radical mastectomy	4,854 (16.4)	4,173 (16.9)	22,241 (12.4)	3,690 (14.1)
Subcutaneous mastectomy	454 (1.5)	306 (1.2)	2,856 (1.6)	575 (2.2)
Total (simple) mastectomy	6,822 (23.1)	4,914 (20.0)	38,660 (21.6)	6,942 (26.5)
Unknown	<11 (<0.1)	32 (0.1)	196 (0.1)	41 (0.2)
Bone metastasis at diagnosis	834 (2.8)	1,063 (4.3)	5,012 (2.8)	616 (2.4)

Results are n (%), except for age at diagnosis, which is median (interquartile range).
* Including common law marriage.
† Includes never married, separated, divorced, and widowed.
‡ Includes no specifics, any Medicaid, and probably Medicare.
HER2–, HER2-negative; HER2+, HER2-positive; HR+, hormone receptor-positive.

Figure 1: Impact of variables on hazard predictions



SES, socioeconomic status.

Figure 2: Racial–ethnic disparities in survival

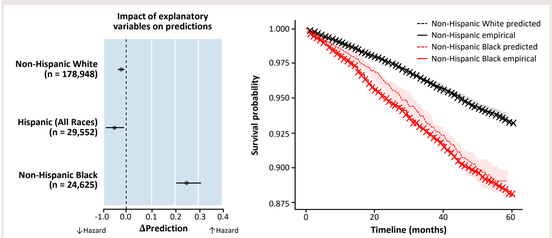
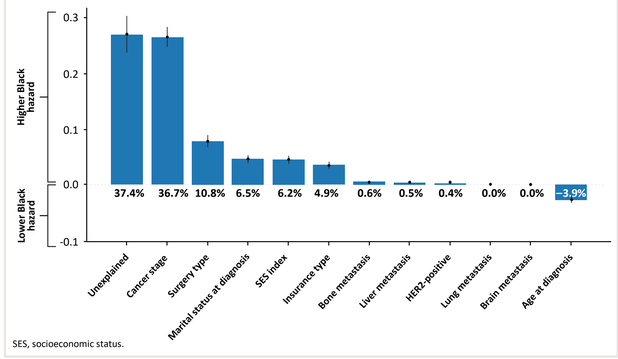


Figure 3: Contributions to predicted survival differences between Non-Hispanic Black and White patients



SES, socioeconomic status.

CONCLUSIONS

- An ML approach was able to accurately predict 5-year cancer-specific survival and characterize how variables contributed to racial–ethnic survival disparities.
- The biggest known driver of disparity between survival outcomes for Non-Hispanic Black and Non-Hispanic White patients with HR+ breast cancer was identified as being cancer staging at diagnosis.
- This suggests that efforts to detect cancer earlier among Non-Hispanic Black women may help to narrow the gap in survival.

REFERENCES

1. Hardy D & Du DY. *J Racial Ethn Health Disparities* 2021; **8**:990–1001;
2. Tammemagi CM. *Curr Opin Obstet Gynecol* 2007; **19**:31–36;
3. Hirschman J, *et al*. *Cancer Causes Control* 2007; **18**:323–333;
4. Ward E, *et al*. *CA Cancer J Clin* 2004; **54**:78–93;
5. Reeder-Hayes KE, *et al*. *Cancer* 2019; **125**:3985–3992;
6. Poteat TC, *et al*. *Cancer* 2021; **127**:3514–3522;
7. Moncada-Torres A, *et al*. *Sci Rep* 2021; **11**:6968;
8. Jansen T, *et al*. Digital Personalized Health and Medicine, IOS Press 2020; 307–311.

ACKNOWLEDGMENTS

Support for third-party writing assistance for this poster, furnished by Katie Wilson, PhD, of Health Interactions, was provided by F. Hoffmann-La Roche Ltd.

CONFLICTS OF INTEREST

All authors are employees of Genentech, Inc. and hold stocks/shares in F. Hoffmann-La Roche Ltd. All authors have received research support in the form of third-party writing assistance for this poster from F. Hoffmann-La Roche Ltd. This study was sponsored by F. Hoffmann-La Roche Ltd.

Poster

https://bit.ly/41P8HQT