Title

Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer.

Journal

BMJ health & care informatics, 30(1)

ISSN

2632-1009

Authors

Park, Jung In
Bozkurt, Selen
Park, Jong Won
et al.

Publication Date

2023

DOI

10.1136/bmjhci-2022-100666

Peer reviewed

BMJ Health & Care Informatics

Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer

Jung In Park ,[1] Selen Bozkurt,[2] Jong Won Park,[3] Sunmin Lee[4]

To cite: Park JI, Bozkurt S, Park JW, *et al*. Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer. *BMJ Health Care Inform* 2023;**30**:e100666. doi:10.1136/bmjhci-2022-100666

Received 18 August 2022
Accepted 29 December 2022

[1]Sue & Bill Gross School of Nursing, University of California Irvine, Irvine, California, USA
[2]Stanford University, Stanford, California, USA
[3]Yonsei University College of Medicine, Seoul, Seodaemun-gu, Korea (the Republic of)
[4]School of Medicine, University of California Irvine, Irvine, California, USA

Correspondence to
Dr Jung In Park;
junginp@uci.edu

ABSTRACT

Objectives Survival machine learning (ML) has been suggested as a useful approach for forecasting future events, but a growing concern exists that ML models have the potential to cause racial disparities through the data used to train them. This study aims to develop race/ethnicity-specific survival ML models for Hispanic and black women diagnosed with breast cancer to examine whether race/ethnicity-specific ML models outperform the general models trained with all races/ethnicity data.

Methods We used the data from the US National Cancer Institute's Surveillance, Epidemiology and End Results programme registries. We developed the Hispanic-specific and black-specific models and compared them with the general model using the Cox proportional-hazards model, Gradient Boost Tree, survival tree and survival support vector machine.

Results A total of 322 348 female patients who had breast cancer diagnoses between 1 January 2000 and 31 December 2017 were identified. The race/ethnicity-specific models for Hispanic and black women consistently outperformed the general model when predicting the outcomes of specific race/ethnicity.

Discussion Accurately predicting the survival outcome of a patient is critical in determining treatment options and providing appropriate cancer care. The high-performing models developed in this study can contribute to providing individualised oncology care and improving the survival outcome of black and Hispanic women.

Conclusion Predicting the individualised survival outcome of breast cancer can provide the evidence necessary for determining treatment options and high-quality, patient-centred cancer care delivery for under-represented populations. Also, the race/ethnicity-specific ML models can mitigate representation bias and contribute to addressing health disparities.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Survival machine learning allows healthcare professionals to identify patients at high risk, but models trained with data poorly representative of minority groups, they may exacerbate health disparities. To date, no study developed race/ethnicity-specific survival machine learning models for Hispanic and black women diagnosed with breast cancer.

WHAT THIS STUDY ADDS

⇒ The race/ethnicity-specific survival machine learning models outperformed the general models trained with all races/ethnicity when predicting the outcomes of specific races/ethnicity.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Predicting the individualised survival outcome of breast cancer can provide the evidence necessary for determining treatment options and high-quality, patient-centred cancer care delivery for underrepresented populations. Also, the race/ethnicity-specific machine learning models can mitigate representation bias and contribute to addressing health disparities.

INTRODUCTION

Breast cancer is the second-leading cause of cancer-related deaths in women in the USA, and it affects every ethnic group of women in the USA.[1 2] However, there are racial and ethnic divides in cancer survival. Breast cancer is the most prevalent reason for cancer-related death in Hispanic women in the USA.[3] Also, minority women, especially black women, have a higher mortality rate (26.8 per 100 000 women) even though white women (18.8 per 100 000 women) have higher cancer incidence.[2 4 5] These facts indicate that the cancer survival rates need to be improved among Hispanic and black women, and various features contributing to breast cancer mortality should be understood to provide tailored intervention for enhanced survival.

Unlike traditional survival models that use a standard statistical method, survival machine learning (ML) has been suggested as a useful approach for learning the patterns from high-dimensional data and complex feature interactions for forecasting future events.[6] This approach allows healthcare professionals to identify patients at high risk or predict those

who need increased utilisation of healthcare services to proactively support and provide interventions necessary for the patients.[7] However, a growing concern exists that ML models have the potential to cause racial disparities through the data used to train them.[8] The ML model trained with the data representing general population would not contain sufficient number of participants from the minority population and is biased, resulting in inaccurate predictions for the minority group even if the overall accuracy is high.[9] If the ML models trained with data poorly representative of minority groups are used in healthcare, they may exacerbate health disparities.[10] To address such harmful effects, it is recommended to train an ML model with data that resemble the population that the model is intended to use.[11 12] To the best of our knowledge, no study developed race/ethnicity-specific survival ML models for Hispanic and black women diagnosed with breast cancer.

Therefore, there is a need for race/ethnicity-specific survival ML models trained with the underrepresented populations to examine the feasibility of race/ethnicity-specific ML models that may outperform the general model trained with all races/ethnicity. Accurate prediction of the individualised outcome will enable tailored healthcare delivery and a better outcome for the underrepresented populations. This study aims to develop race/ethnicity-specific survival ML models for Hispanic and black women diagnosed with breast cancer to examine whether race/ethnicity-specific ML models outperform the models trained with the general population data when predicting the survival of Hispanic and black women diagnosed with breast cancer.

METHODS

Data source

We used the data from the US National Cancer Institute's population-based Surveillance, Epidemiology and End Results (SEER) programme registries. The SEER programme currently collects and publishes cancer incidence and survival data in the USA from population-based cancer registries in 22 geographical areas, representing approximately 48% of the US population.[13] The SEER data are considered the gold standard for data quality among cancer registries in the USA and globally.[14] We selected adult female patients' data (18 or older) from SEER who had breast cancer diagnoses between 1 January 2000 and 31 December 2017. Also, we selected California as the geographical location for the diverse characteristics of the patient population. The Hispanic population included all races, and the black population was non-Hispanic. Figure 1 shows the flow chart of data collection.

Predictor and outcome variables

The predictor variables included age at cancer diagnosis, marital status at diagnosis, first malignant primary tumour indicator, the sequence number of tumours, primary site, histology, the total number of in situ/malignant tumours,

Figure 1 Flow chart of data collection.

SEER summary stage, derived stage, grade, regional lymph nodes examined, regional lymph nodes positive, oestrogen receptor status, progesterone receptor status, chemotherapy, radiation, sequence of radiation and surgery performed, reason no cancer-directed surgery and sequence of systemic therapy and surgical procedures. Vital status was recorded as alive/dead at the time of the cut-off date (31 December 2017). The sequence number of tumours describes the sequence of all reportable tumours that occurred over a patient's lifetime.

The outcome variable was the survival months of a patient.

Data preprocessing and preparation

Before training the survival models, we preprocessed the predictor variables to enhance the ML modelling performance. Rows containing missing values were dropped. All the categorical features were reencoded using a one-hot-encoding scheme where each new column represented a single category. We applied variance filtering (with the threshold of 0.01) to drop the features that were near-constant or had low variance. Thus, a feature containing outliers would appear as a low-variance column and be filtered out. Once the preprocessing was completed, the final dataset was exported into a new flat file for the training. To train an ML model for survival analysis, the 'survival months' variable was used as the target for the training. 'Vital status' was used for the event.

We took several steps for data preparation to develop race/ethnicity-specific models for the Hispanic and black populations and compare them with the general model that included all races/ethnicity. Figure 2 shows the process of data preparation for model development.

First, we split the full dataset into a training set (T_{all}) for model development and a test set (E_{all}) for evaluation with a 7:3 ratio to randomly sample the populations. Each set was used to sample the populations for model

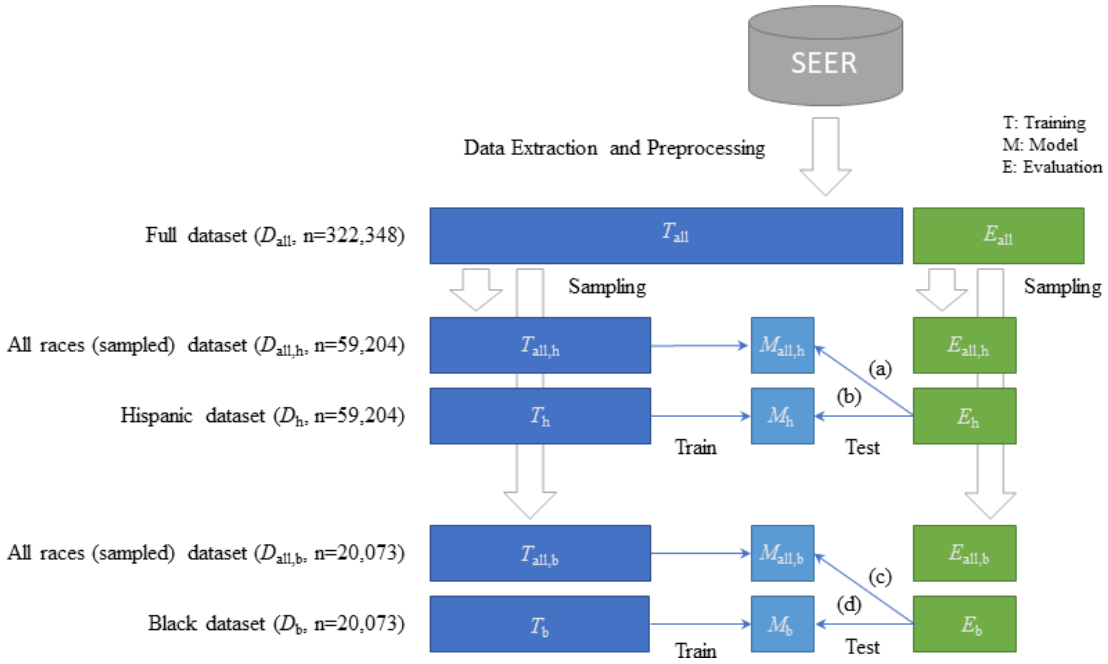

Figure 2 Data preparation for model development. SEER, Surveillance, Epidemiology and End Results.

development randomly. The randomly sampled population sets maintained the original ratio of each race/ethnicity in the full dataset. Second, we extracted the Hispanic population from the original training set, T_{all} (T_h) to train the Hispanic-specific model (M_h). We also extracted the Hispanic population from the original test set, E_{all} (E_h) to test the model, M_h. Then, we randomly sampled the populations from the original training set (T_{all}) that included all races/ethnicity ($T_{all,h}$), to match the exact number of samples used for the Hispanic-specific model training. We also randomly sampled the populations from the original test set (E_{all}) that included all races/ethnicity ($E_{all,h}$), to match the exact number of samples used for the Hispanic-specific model testing. $T_{all,h}$ was used to develop a model $M_{all,h}$. Then, the performance of the models $M_{all,h}$ (a) and M_h (b) were compared with the same test set, E_h. Third, we repeated the process of Hispanic-specific model development for Black-specific model development.

We extracted the black population from the original training set, T_{all} (T_b) to train the black-specific model (M_b). We also extracted the black population from the original test set, E_{all} (E_b) to test the model, M_b. Then, we randomly sampled the populations from the original training set (T_{all}) that included all races/ethnicity ($T_{all,b}$), to match the exact number of samples used for the black-specific model training. We also randomly sampled the populations from the original test set (E_{all}) that included all races/ethnicity ($E_{all,b}$), to match the exact number of samples used for the black-specific model testing. $T_{all,b}$ was used to develop a model $M_{all,b}$. Then, the performance of the models $M_{all,b}$ (c) and M_b (d) were compared with the same test set, E_b.

Race/ethnicity-specific models

For the survival ML modelling, we developed and compared four models: Cox proportional-hazards (PH) model (CoxPH), Gradient Boost Tree (GBT), survival

Table 1 Description of survival machine learning models

Model	Description
Cox PH	A standard survival model looking at the effects of a patient's covariates on the risk of death.[26] It is a multivariate regression model for survival analysis.[27]
GBT	An ensemble learning method that sequentially combines the outputs from individual decision trees, so each new tree can predict and correct the errors of the previous tree.[28] It uses Gradient-boosted Cox proportional hazard loss with regression trees as base learner.[29]
ST	A model that splits the covariate space into smaller nodes containing observations with homogeneous survival outcomes.[30] It is a tree-based method for censored survival data.[31]
SSVM	An extension of the standard SVM to maximise the concordance index (*C-index*) and account for complex, non-linear relationships between features and survival.[32] It is an efficient way of training a kernel SVM.[33]

GBT, Gradient Boost Tree; PH, proportional hazard; SSVM, survival support vector machine; ST, survival tree.

tree (ST) and survival support cector machine (SSVM). The description of each model is shown in table 1.

Each model's performance was evaluated using the C-index. The C-index is a standard way of measuring the performance of survival models. It can be viewed as the fraction of all pairs of patients predicted to have correct orders over the total number of possible evaluation pairs.[15]

For each race/ethnicity, we trained and compared two different models based on the two datasets mentioned above—one with a specific race/ethnicity and the other one with all races/ethnicity. Our hypothesis was that the model trained with specific race/ethnicity would outperform the general model trained with all races/ethnicity when predicting the breast cancer survival of a specific race/ethnicity.

RESULTS

Sample characteristics

A total of 322 348 female patients who had breast cancer diagnoses between 1 January 2000 and 31 December 2017 were identified. Among them, the number of Hispanic patients was 59 204 (18.4%), and black was 20 073 (6.2%). Table 2 shows the detailed characteristics of the study sample, Hispanic, black and all races/ethnicity.

Compared with all races/ethnicity (15.2%) and Hispanic (14.9%) populations, more black population was dead (24.4%). Hispanic population's survival months (mean: 80.6, median: 67.0) were lower compared with all races/ethnicity (90.4, 79.0) and black (82.9, 69.0) populations. Hispanic population was younger (mean:55.3, median 54.0), compared with all races/ethnicity (59.1, 59.0) and black (57.8, 57.0) populations. Black (36.6%) and Hispanic (39.2%) population had higher percentage of poorly differentiated grade III cancer, compared with all races/ethnic (32.7%) groups. Black population had lower percentages of positive oestrogen receptor status (65.6%), compared with all races/ethnicity (77.4%) and Hispanic (73.5%) populations. Also, black population had lower percentages of positive progesterone receptor status (52.1%), compared with all races/ethnicity (65.6%) and Hispanic (62.3%) populations.

Lower percentages of Hispanic (51.0%) and black (50.0%) populations had chemotherapy compared with all races/ethnicity (57.4%). Higher percentages of black (57.3%) and Hispanic (55.6%) populations had no radiation and/or cancer-directed surgery, compared with all races/ethnicity (52.3%). Higher percentages of overall (46.4%) and Hispanic (43.4%) populations had radiation after surgery than Black (41.5%) populations.

Figure 3 shows the Kaplan-Meier curves of the Hispanic, black and all races/ethnic groups. All races/ethnic groups had the better survival than the Hispanic and black groups.

Data preprocessing and preparation

After data preprocessing and cleaning, the final dataset for analysis contained 260 variables. Values in 'Derived stages' variable were grouped into '0', 'I', 'II', 'III', 'IV' and 'unknown'. 'Regional nodes examined' and 'Regional nodes positive' were integer variables which contains both numeric and encoded values (90+). Numeric values were categorised (ie, 0–9, 11–19, …, 40+), while encoded values were mapped to 'oher'.

Model development

We extracted 59 204 Hispanic populations for each training set for Hispanic-specific model (T_h) and a comparison model with all races/ethnicity ($T_{all,h}$). Also, we extracted 20 073 black populations for each training set for black-specific model (T_b) and a comparison model with all races/ethnicity ($T_{all,b}$). Once data were prepared, we applied variance filtering and dropped the features that had low variance. After filtering, the number of features we had for the T_h was 72, and for the $T_{all,h}$ was 71 for Hispanic-specific model training, and the number of features we had for the T_b and $T_{all,b}$ was 72 for the black-specific model training.

During the training, both training sets (race-specific and all races/ethnicity) were further split into actual training set and validation set during a cross-validation phase when parameter tuning was necessary (GBT and ST models). We used random search method to find the most optimal parameters for each survival analysis model. We used 20 iterations and 5-fold cross validation was used for all cases for each training. We used scikit-survival package (V.0.17.1) for the modelling (CoxPHSurvivalAnalysis class for CoxPH, Gradient Boosting Survival Analysis class for GBT, SurvivalTree class for ST and FastKernelSurvivalSVM class for SSVM), scikit-learn (V.1.0.2) for the feature selection (VarianceThreshold), hyperopt (V.0.2.7) for the hyperparameter search, and pandas (V.1.4.1) for general data preprocessing and preparation.

Model evaluations

The model evaluation results are shown in table 3 and figure 4 where we compared different combinations of modelling methods and input training/test sets.

Hispanic-specific model (M_h) and all races/ethnicity model ($M_{all,h}$) were evaluated using the same the test set (E_h). Hispanic-specific model (M_h) outperformed all races/ethnicity model ($M_{all,\ h}$) in three out of four approaches, which were Cox PH (0.832 vs 0.828), ST (0.772 vs 0.763) and SSVM (0.834 vs 0.790). The GBT model showed the same c-index score (0.813) for both models.

Black-specific model (M_b) and all races/ethnicity model ($M_{all,b}$) were evaluated using the same the test set (E_b). Black-specific model (M_b) outperformed all races/ethnicity model ($M_{all,b}$) in all four approaches, Cox PH (0.823 vs 0.821), GBT (0.808 vs 0.803), ST (0.804 vs 0.801) and SSVM (0.824 vs 0.786). In both race/

Table 2 Sample characteristics

Non-Hispanic (NH) white	199 913 (62.0)		
Hispanic (all races)	**59 204 (18.4)**		
NH Asian or Pacific Islander	41 811 (13.0)		
NH black	**20 073 (6.2)**		
NH American Indian/Alaska Native	1347 (0.4)		
Total	322 348		
	All races/ethnicity (N=322 348)	**Hispanic (N=59 204)**	**Black (N=20 073)**
Vital status (n, %)			
Alive	273 455 (84.8)	50 382 (85.1)	15 175 (75.6)
Dead	48 893 (15.2)	8822 (14.9)	4898 (24.4)
Survival months			
Min, Max	0.0, 227.0	0.0, 227.0	0.0, 227.0
Mean	90.4	80.6	82.9
Median	79.0	67.0	69.0
SD	60.9	58.7	59.8
Age			
Min, Max	19.0, 100.0	19.0, 99.0	19.0, 100.0
Mean	59.1	55.3	57.8
Median	59.0	54.0	57.0
SD	13.0	12.9	13.0
First malignant primary indicator (n, %)			
Y	278 117 (86.3)	53 174 (89.8)	17 240 (85.9)
N	44 231 (13.7)	6030 (10.2)	2833 (14.1)
Sequence no of tumours (n, %)			
One primary only	236 182 (73.3)	46 956 (79.3)	14 650 (73.0)
Second of two or more primaries	44 226 (13.7)	6233 (10.5)	2812 (14.0)
First of two or more primaries	34 893 (10.8)	5328 (9.0)	2152 (10.7)
Third of three or more primaries	6008 (1.9)	603 (1.0)	397 (2.0)
Fourth of four or more primaries	878 (0.3)	75 (0.1)	51 (0.3)
Other	161 (0.0)	9 (0.0)	11 (0.1)
Histology (n, %)			
Ductal and lobular neoplasms	306 634 (95.1)	56 417 (95.3)	18 904 (94.2)
Cystic, mucinous and serous neoplasms	5800 (1.8)	1001 (1.7)	411 (2.0)
Adenomas and adenocarcinomas	5063 (1.6)	768 (1.3)	294 (1.5)
Epithelial neoplasms, NOS	1343 (0.4)	279 (0.5)	139 (0.7)
Complex epithelial neoplasms	1311 (0.4)	258 (0.4)	154 (0.8)
Adnexal and skin appendage neoplasms	807 (0.3)	136 (0.2)	65 (0.3)
Squamous cell neoplasms	573 (0.2)	116 (0.2)	52 (0.3)
Fibroepithelial neoplasms	385 (0.1)	141 (0.2)	19 (0.1)
Other	432 (0.1)	88 (0.1)	35 (0.2)
Total no of in situ/malignant tumours (n, %)			
1	239 762 (74.4)	47 536 (80.3)	14 880 (74.1)
2	67 146 (20.8)	10 046 (17.0)	4224 (21.0)
3	12 622 (3.9)	1387 (2.3)	802 (4.0)
4	2276 (0.7)	206 (0.3)	141 (0.7)

Continued

Table 2 Continued

5+	542 (0.2)	29 (0.0)	26 (0.1)
Summary stage (n, %)			
Localised	205 612 (63.8)	34 061 (57.5)	11 398 (56.8)
Regional	102 914 (31.9)	22 292 (37.7)	7291 (36.3)
Distant	13 822 (4.3)	2851 (4.8)	1384 (6.9)
Grade (n, %)			
Moderately differentiated; grade II	138 937 (43.1)	24 491 (41.4)	9346 (46.6)
Poorly differentiated; grade III	105 361 (32.7)	23 237 (39.2)	7350 (36.6)
Well differentiated; grade I	73 815 (22.9)	10 607 (17.9)	3012 (15.0)
Undifferentiated; anaplastic; grade IV	4235 (1.3)	869 (1.5)	365 (1.8)
Regional lymph nodes examined (n, %)			
0–9	233 642 (72.5)	39 268 (66.3)	13 679 (68.1)
10–19	61 685 (19.1)	13 312 (22.5)	4470 (22.3)
20–29	17 330 (5.4)	4167 (7.0)	1163 (5.8)
Other	6275 (1.9)	1547 (2.6)	549 (2.7)
30–39	2837 (0.9)	740 (1.2)	176 (0.9)
40+	579 (0.2)	170 (0.3)	36 (0.2)
Oestrogen receptor status (n, %)			
Positive	249 656 (77.4)	43 494 (73.5)	13 168 (65.6)
Negative	56 876 (17.6)	12 575 (21.2)	5941 (29.6)
Borderline/unknown	15 071 (4.7)	2925 (4.9)	915 (4.6)
N/A	745 (0.2)	210 (0.4)	49 (0.2)
Progesterone receptor status (n, %)			
Positive	210 985 (65.5)	36 867 (62.3)	10 463 (52.1)
Negative	90 699 (28.1)	18 369 (31.0)	8224 (41.0)
Borderline/unknown	19 919 (6.2)	3758 (6.3)	1337 (6.7)
N/A	745 (0.2)	210 (0.4)	49 (0.2)
Chemotherapy (n, %)			
Y	185 019 (57.4)	30 197 (51.0)	10 038 (50.0)
N	137 329 (42.6)	29 007 (49.0)	10 035 (50.0)
Radiation (n, %)			
None/unknown	150 350 (46.6)	29 136 (49.2)	9965 (49.6)
Beam radiation	147 114 (45.6)	25 651 (43.3)	8429 (42.0)
Recommended, unknown if administered	10 290 (3.2)	2532 (4.3)	922 (4.6)
Refused	5901 (1.8)	796 (1.3)	403 (2.0)
Radioactive implants	5897 (1.8)	645 (1.1)	191 (1.0)
Radiation, NOS method or source not specified	2415 (0.7)	367 (0.6)	150 (0.7)
Other	381 (0.1)	77 (0.1)	13 (0.1)
Sequence of radiation and surgery performed (n, %)			
No radiation and/or cancer-directed surgery	168 694 (52.3)	32 906 (55.6)	11 505 (57.3)
Radiation after surgery	149 615 (46.4)	25 718 (43.4)	8340 (41.5)
Intraoperative radiation	1839 (0.6)	173 (0.3)	68 (0.3)
Radiation prior to surgery	728 (0.2)	168 (0.3)	80 (0.4)
Radiation before and after surgery	673 (0.2)	146 (0.2)	40 (0.2)
Intraoperative rad with other rad before/after surgery	560 (0.2)	53 (0.1)	28 (0.1)
Other	239 (0.1)	40 (0.1)	12 (0.1)

Continued

Table 2 Continued

Reason no cancer-directed surgery (n, %)			
Surgery performed	305 404 (94.7)	55 371 (93.5)	18 308 (91.2)
Not recommended	13 032 (4.0)	2952 (5.0)	1356 (6.8)
Recommended, unknown if performed	1899 (0.6)	568 (1.0)	217 (1.1)
Recommended but not performed, patient refused	1364 (0.4)	211 (0.4)	146 (0.7)
Not recommended, contraindicated due to other	333 (0.1)	53 (0.1)	22 (0.1)
conditions, autopsy only	285 (0.1)	44 (0.1)	22 (0.1)
Other	31 (0.0)	5 (0.0)	2 (0.0)

N/A, not available.

ethnicity-specific models, Cox PH showed the highest c-index score followed by GBT, SSVM and ST.

DISCUSSION

Accurately predicting the survival outcome of a patient is critical in determining treatment options and providing appropriate cancer care. The ML approaches provide a robust way of predicting health outcomes using large data points with complex feature interactions. However, current ML models are often built with all races/ethnicity data, having the potential to have representation bias, and not tailored to each minority group. To date, race/ethnicity-specific survival ML models predicting the outcomes of the black and Hispanic women diagnosed with breast cancer are lacking. This study developed and evaluated race/ethnicity-specific survival ML models for black and Hispanic women with breast cancer and compared with the general population model. The high performing ML models developed in this study will be able to contribute to providing individualised oncology care and improving the survival outcome of specific populations, the black and Hispanic women. Also, it is a strength of our model that we used the patient data from more than 322 348 women in a large, population-based dataset from 2000 to 2017, including 59 204 (18.4%) Hispanic women and 20 073 (6.2%) Black women.

The sample population in this study showed that the black population had the highest death rate followed by the Hispanic and all races/ethnicity, supporting the findings from other literature.[4 5] Also, the survival months for the black and Hispanic groups were low and they were younger compared with all races/ethnicity. It is congruent with the literature that young black women have higher breast cancer mortality than young white women,[16 17] and the Latinas have the higher rates of more advanced cancer than non-Hispanic Whites.[18] Also, breast cancer is more aggressive in younger women than older premenopausal women.[19] Our study sample also showed that the Hispanic and black populations had higher percentage of poorly differentiated grade III cancer than overall populations. Poorly differentiated tumours lack normal features, tend to grow and spread faster and have a worse prognosis[20]; and these tumours expressed lower levels of oestrogen receptor.[21] Our study sample showed likewise that Hispanic and black populations showed the lower percentage of oestrogen receptor positive status and progesterone receptor positive status than overall population. Studies have shown that young age breast cancer has more advanced stage at presentation, more grades and higher oestrogen receptor negativity.[22]

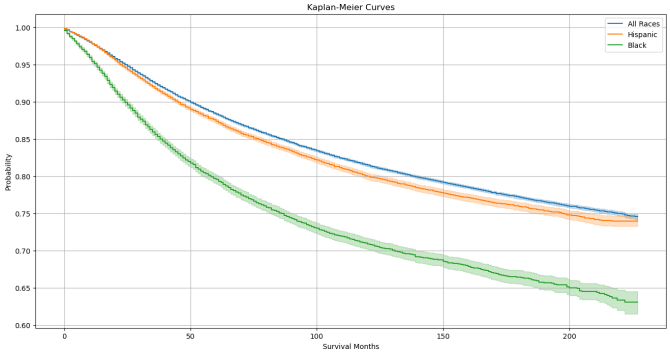

Figure 3 Kalplan-Meier survival curves.

The result also showed that lower percentages of Hispanic and black populations had chemotherapy. Existing literature has shown that African American and Hispanic patients tend to experience diagnostic and treatment delays, which were related to worse survival outcomes.[23 24] Perhaps lower percentages of Hispanic and black patients receiving chemotherapy were associated with the fewer survival months of the Hispanic and black populations in this study.

After the race/ethnicity-specific model development and evaluation, we observed that the general models trained with all races/ethnicity did not perform well when tested with specific races/ethnicity. That is, the race/ethnicity-specific survival ML models developed in this study consistently outperformed the general models when predicting the outcomes of specific race/ethnicity, addressing bias in ML. Especially, black and Hispanic-specific survival ML models using the Cox PH approach showed the best performance among the four ML models tested, showing that this model outperformed the other models in predicting the survival of specific race/ethnicity. Also, the ST model performance showed the

Table 3 Model performance comparison using c-index

	Hispanic-specific model	All races/ethnicity model	Black-specific model	All races/ethnicity model
Model	M_h	$M_{all,h}$	M_b	$M_{all,b}$
Test	E_h	E_h	E_b	E_b
Cox PH	0.832	0.828	0.823	0.821
GBT	0.813	0.813	0.808	0.803
ST	0.772	0.763	0.804	0.801
SSVM	0.834	0.790	0.824	0.786

GBT, Gradient Boost Tree; PH, proportional hazard; SSVM, survival support vector machine; ST, survival tree.

highest difference between the race/ethnicity-specific model and the general model. This indicates that the ST model tends to overfit to a specific race/ethnicity compared with the other models. Our study demonstrated that a tailored ML model for each race/ethnicity is needed to better predict the patient survival than the general ML model using all races/ethnicity. By accurately forecasting a patient's survival, healthcare professionals will be able to guide individualised treatment decisions and provide tailored interventions for the well-being of a cancer survivor.

It is worth noting that although the performance of the general model is not low, it was trained with the general population with an imbalanced portion of the underrepresented population, including the Hispanic and black populations. It was still meaningful to examine the feasibility of race/ethnicity-specific models since it is recommended to train an ML model with data resembling the people the model is intended to use to mitigate representation bias. Although the performance difference between the models was sometimes marginal depending on the algorithms, our race/ethnicity-specific models consistently outperformed the general model. It shows the potential to accurately predict individualised patient outcomes for quality care delivery for underrepresented populations and lead to alleviating health disparities.

There are several limitations to this study. The SEER database only includes the first course of treatment and do not have information on adjuvant therapy.[25] This causes difficulties comparing the outcomes of the treatment sequence. To overcome this limitation, a comprehensive database that has more information on cancer treatment can be used as a future work to provide additional insights on the impact of treatment sequence. Also, the dataset did not include the human epidermal growth factor 2 receptor status, which is a critical tumour marker for breast cancer prognosis. The variable was missing because it was collected from 2010, but our data were dated from 2000. Incorporating this variable in the modelling will be needed in future work to provide more accurate predictions for patient outcomes.

CONCLUSION

This study has developed and evaluated accurate race/ethnicity-specific survival ML models for black and Hispanic women diagnosed with breast cancer. Predicting the individualised survival outcome of breast cancer can

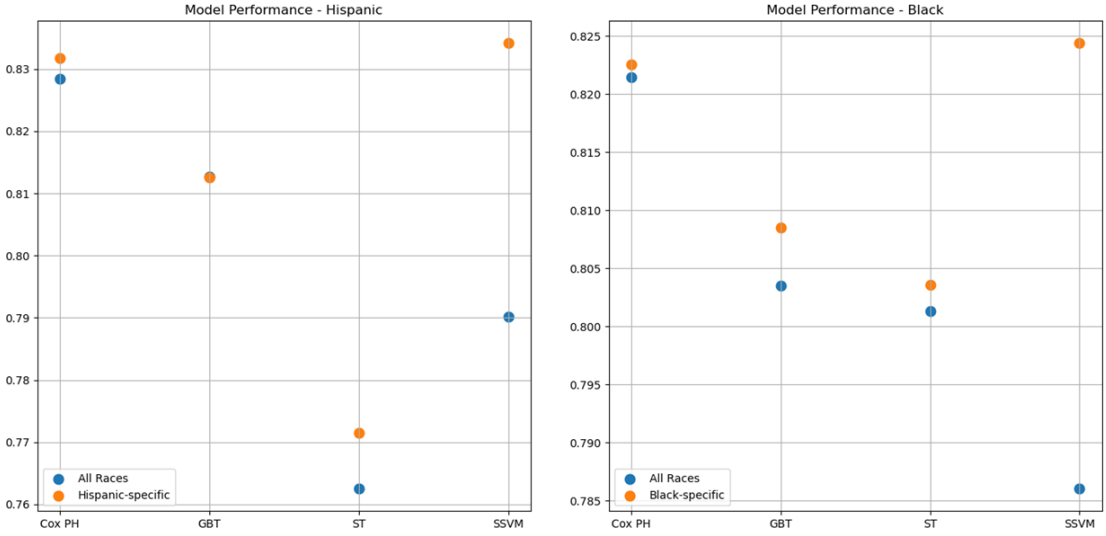

Figure 4 Race/ethnicity-specific model performance comparison using C-index. GBT, Gradient Boost Tree; PH, proportional hazards; SSVM, survival support vector machine; ST, survival tree.

provide the evidence necessary for determining treatment options and high-quality, patient-centred cancer care delivery for underrepresented populations. Also, the race/ethnicity-specific ML models can mitigate representation bias and contribute to addressing health disparities.

Contributors All the authors contributed to the design of the work and the final approval of the submission. JIP worked on the data acquisition, analysis and interpretation of the data, and acted as guarantor. SB contributed to the data analysis. JWP and SL contributed to the interpretation of data for the work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Since the data were fully deidentified, this study was not considered human subject research by the Institutional Review Board at the University, and no informed consent was required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. We used the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) Program data. It provides information on cancer statistics in an effort to reduce the cancer burden among the US population.

ORCID iD

Jung In Park http://orcid.org/0000-0002-1771-7361

REFERENCES

1 Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics, 2021. *CA Cancer J Clin* 2021;71:7–33.
2 Yedjou CG, Sims JN, Miele L. Health and racial disparity in breast cancer. *Breast cancer metastasis and drug resistance* 2019;1152:31–49.
3 Power EJ, Chin ML, Haq MM. Breast cancer incidence and risk reduction in the Hispanic population. *Cureus* 2018;10:e2235.
4 Jemal A, Ward EM, Johnson CJ, *et al*. Annual report to the nation on the status of cancer, 1975–2014, featuring survival. *J Natl Cancer Inst* 2017;109.
5 Copeland G, Green D, Firth R. Cancer in North America: 2011–2015 volume one: combined cancer incidence for the United States, Canada and North America. Springfield North American Association of Central Cancer Registries; 2018.
6 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.
7 Bates DW, Saria S, Ohno-Machado L, *et al*. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff* 2014;33:1123–31.
8 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
9 Saria S, Subbaswamy A. Tutorial: safe and reliable machine learning. *arXiv preprint* 2019.
10 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15:e1002689.
11 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
12 Wiens J, Saria S, Sendak M, *et al*. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337–40.
13 SEER, National Cancer Institute. SEER research plus data description cases diagnosed in 1975-2017, 2020. Available: https://seer.cancer.gov/data-software/documentation/seerstat/nov2019/TextData.FileDescription.pdf
14 Duggan MA, Anderson WF, Altekruse S, *et al*. The surveillance, epidemiology, and end results (SEER) program and pathology: toward strengthening the critical relationship. *Am J Surg Pathol* 2016;40:e94.
15 Chen Y, Jia Z, Mercola D, *et al*. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput Math Methods Med* 2013;2013:1–8.
16 Shavers VL, Harlan LC, Stevens JL. Racial/Ethnic variation in clinical presentation, treatment, and survival among breast cancer patients under age 35. *Cancer* 2003;97:134–47.
17 Althuis MD, Brogan DD, Coates RJ, *et al*. Breast cancers among very young premenopausal women (United States). *Cancer Causes Control* 2003;14:151–60.
18 Yanez B, Thompson EH, Stanton AL. Quality of life among Latina breast cancer patients: a systematic review of the literature. *J Cancer Surviv* 2011;5:191–207.
19 Gonzalez-Angulo AM, Broglio K, Kau S-W, *et al*. Women age < or = 35 years with primary breast carcinoma: disease features at presentation. *Cancer* 2005;103:2466–72.
20 American Cancer Society. Understanding your pathology report: breast cancer, 2020. Available: https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/breast-pathology/breast-cancer-pathology.html
21 Scimeca M, Antonacci C, Colombo D, *et al*. Emerging prognostic markers related to mesenchymal characteristics of poorly differentiated breast cancers. *Tumour Biol* 2016;37:5427–35.
22 Zabicki K, Colbert JA, Dominguez FJ, *et al*. Breast cancer diagnosis in women < or = 40 versus 50 to 60 years: increasing size and stage disparity compared with older women over time. *Ann Surg Oncol* 2006;13:1072–7.
23 Gwyn K, Bondy ML, Cohen DS, *et al*. Racial differences in diagnosis, treatment, and clinical delays in a population-based study of patients with newly diagnosed breast carcinoma. *Cancer* 2004;100:1595–604.
24 Fedewa SA, Ward EM, Stewart AK, *et al*. Delays in adjuvant chemotherapy treatment among patients with breast cancer are more likely in African American and Hispanic populations: a national cohort study 2004-2006. *J Clin Oncol* 2010;28:4135–41.
25 Yu JB, Gross CP, Wilson LD, *et al*. NCI SEER public-use data: applications and limitations in oncology research. *Oncology* 2009;23:288.
26 Katzman JL, Shaham U, Cloninger A, *et al*. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018;18:1–2.
27 Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* 1989;84:1074–8.
28 Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;38:367–78.
29 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 2001;29:1189–232.
30 Bertsimas D, Dunn J, Gibson E, *et al*. Optimal survival trees. *Mach Learn* 2022;111:2951–3023.
31 Leblanc M, Crowley J. Survival trees by Goodness of split. *J Am Stat Assoc* 1993;88:457–67.
32 Van Belle V, Pelckmans K, Van Huffel S, *et al*. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med* 2011;53:107–18.
33 Pölsterl S, Navab N, Katouzian A. An efficient training algorithm for kernel survival support vector machines. *arXiv preprint* 2016.