Cancer Epidemiology, Biomarkers & Prevention

Differences in Breast Cancer Survival by Molecular Subtypes in the United States

Nadia Howlader[1], Kathleen A. Cronin[1], Allison W. Kurian[2], and Rebecca Andridge[3]

Abstract

Background: Although incidence rates of breast cancer molecular subtypes are well documented, effects of molecular subtypes on breast cancer–specific survival using the largest population coverage to date are unknown in the U.S. population.

Methods: Using Surveillance, Epidemiology and End Results cancer registry data, we assessed survival after breast cancer diagnosis among women diagnosed during 2010 to 2013 and followed through December 31, 2014. Breast cancer molecular subtypes defined by joint hormone receptor [HR, estrogen receptor (ER) and/or progesterone receptor (PR)] and HER2 status were assessed. Multiple imputation was used to fill in missing receptor status. Four-year breast cancer–specific survival per molecular subtypes and clinical/demographic factors were calculated. A Cox proportional hazards model was used to evaluate survival while controlling for clinical and demographic factors.

Results: The best survival pattern was observed among women with $HR^+/HER2^-$ subtype (survival rate of 92.5% at 4 years), followed by $HR^+/HER2^+$ (90.3%), $HR^-/HER2^+$ (82.7%), and finally worst survival for triple-negative subtype (77.0%). Notably, failing to impute cases with missing receptor status leads to overestimation of survival because those with missing receptor status tend to have worse prognostic features. Survival differed substantially by stage at diagnosis. Among *de novo* stage IV disease, women with $HR^+/HER2^+$ subtype experienced better survival than those with $HR^+/HER2^-$ subtype (45.5% vs. 35.9%), even after controlling for other factors.

Conclusions: Divergence of survival curves in stage IV $HR^+/HER2^+$ versus $HR^+/HER2^-$ subtype is likely attributable to major advances in HER2-targeted treatment.

Impact: Contrary to conventional thought, $HR^+/HER2^+$ subtype experienced better survival than $HR^+/HER2^-$ in advanced-stage disease. *Cancer Epidemiol Biomarkers Prev; 27(6); 619–26. ©2018 AACR.*

Introduction

Breast cancer molecular subtypes have been defined based on gene expression profiling (1). The major subtypes of breast cancer are approximated by the joint expression of three tumor markers: estrogen receptor (ER), progesterone receptor [PR, where ER and PR status are jointly assessed as hormone receptor (HR) status], and HER2 status. The four main molecular subtypes approximated by joint HR/HER2 status are: $HR^+/HER2^-$ (i.e., approximating Luminal A subtype), $HR^+/HER2^+$ (Luminal B), $HR^-/HER2^+$ (HER2-enriched), and $HR^-/HER2^-$ (triple-negative; refs. 2–4). Recently, these three tumor markers have become part of routine data collection by the population-based Surveillance, Epidemiology, and End Results (SEER) cancer registries covering about 30% of the U.S. population (3, 5, 6).

Important differences in demographic and clinical characteristics in occurrence of molecular subtypes among women have been described using U.S. population–based cancer registry data (3, 4, 7). In contrast, breast cancer–specific survival for each subtype is poorly documented at the national level. The use of Herceptin and other targeted therapies for HER2-positive breast cancer has been clearly shown to improve survival (8–10). The clinical benefits of these and other treatment advances underscore the need for national population-based data that are specific to subtype. Although some studies have assessed outcomes by molecular subtypes, they were based on relatively small observational studies or confined to a specific geographic region (11–13). Recently, a few studies utilized data from the population-based California Cancer Registry (14–18). However, most of the prior studies examining breast cancer prognosis by subtypes suffered from a large number of women with missing information on receptor status, where missing receptor statuses ranged from 12% to 33% of overall breast cancer cases (3, 11, 17–19). Furthermore, women with missing receptor status tended to be minorities, older, of lower socioeconomic status, had no insurance, or had advanced disease (3). With registry data, we are unable to assume that receptor information is missing completely at random; therefore, failing to account for missing information could bias breast cancer prognosis in the general population.

Our primary aim was to present the first report of nationally representative breast cancer–specific survival estimates by four main molecular subtypes. We provide an assessment of demographic (age, race, ethnicity, socioeconomic status, marital, and insurance status) and clinical (tumor stage, size, nodal status, and Bloom–Richardson grade) differences in breast cancer survival by subtypes using SEER data covering approximately

[1]Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland. [2]Stanford University School of Medicine, Stanford, California. [3]The Ohio State University College of Public Health, Columbus, Ohio.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Nadia Howlader, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892. Phone: 240-276-6891; Fax: 240-277-6591; E-mail: howladern@mail.nih.gov

doi: 10.1158/1055-9965.EPI-17-0627

30% of the U.S. general population. Our second aim was to develop an imputation algorithm to fill in missing receptor status, enabling us to obtain accurate estimates of breast cancer survival in the population.

Materials and Methods

Study population

We identified females with invasive breast cancer diagnosed during 2010–2013 (ICD-O-3 site codes C500-509 excluding histology codes 9050–9055; 9140; 9590–9992). Cases were followed until December 31, 2014 (allowing a maximum follow-up of 59 months). We used data from 17 population-based cancer registries that participate in the SEER program. Data from the Alaska Native Registry were excluded ($n = 273$), since this registry does not provide any additional information on ethnic origin of these cases, i.e., whether they are of Hispanic versus non-Hispanic origin, a distinction necessary for analyses by mutually exclusive race-ethnic subgroups (i.e., non-Hispanic white, non-Hispanic black, non-Hispanic Asian Pacific Islander (API), non-Hispanic American Indian/Alaskan Native, and Hispanic). We began the analysis in 2010 when information on HER2 status was first captured uniformly across all SEER registries (information on ER and PR status in the registries has been collected since 1990). In constructing the study cohort, we applied several exclusion criteria typically used with SEER data for survival analyses: (i) cases diagnosed by autopsy or death certificate ($n = 937$), (ii) alive with no survival time ($n = 466$), (iii) missing cause of death (COD, $n = 584$). We further restricted analysis to cases that were a woman's first or only breast cancer to create a more homogenous group, because a prior cancer diagnosis may affect patient prognosis. The final analytic set consisted of 196,094 females diagnosed with invasive breast cancer.

Study variables

Study variables such as ER, PR, and HER2 status, demographic characteristics (e.g., age, race, ethnicity, year of diagnosis, marital, and insurance status), and tumor stage, size, nodal status, and Bloom–Richardson grade were ascertained across SEER registries using standardized coding rules based on hospital medical records and pathology reports. The four main breast cancer molecular subtypes were derived based on the joint expressions of the three tumor markers explained in detail elsewhere (3). Area-based indicators of patient poverty were derived at the county level from 2010 U.S. Census data and included as tertiles of the observed distribution in the U.S. general population. The proportion of the population in a patient's county of residence who lived below the poverty level was stratified as <10.0% (high socioeconomic status, SES), 10%–19.9% (medium SES), ≥20% (low SES), or unknown. The poverty indicator was used as a surrogate for low, medium, or high SES. Similarly, area-based indicators of urbanicity were derived at the county level from 2010 U.S. Census data and included as the median of the observed distribution in the U.S. general population. The proportion of the population in a patient's county of residence living in urban areas was stratified as <50% (less urban), ≥50% (more urban), or unknown.

Underlying COD

Underlying CODs were ascertained by cancer registries from death certificate codes obtained from the National Center for Health Statistics. To correct for known errors with COD attribution, the SEER program recently developed a special COD variable that maps underlying CODs to the primary cancer diagnosis (20). We used this variable to assign a broad set of CODs to capture deaths due to breast cancer among women with an incident breast cancer diagnosis.

Statistical analysis

Multiple imputation. The three receptor statuses (ER, PR, and HER2) had a considerable amount of missing information. Overall, 16,563 (8.4%) women had a missing HER2 status, 7,965 (4.0%) were missing ER status, and 8,763 (4.5%) were missing PR status. As a result, molecular subtypes could not be derived for 16,996 (8.7%) female breast cancer cases (Table 1). Of these 16,996 cases with missing molecular subtypes, 8,996 (52.9%) women had a known HR status but missing HER2 status, 7,567 (44.5%) women were missing both HR and HER2 statuses, and 433 (2.5%) women had a known HER2 status but missing HR status. In addition, some degree of missing data was present in clinical and demographic variables, which has been documented before (2, 3, 21, 22). For this reason, we used a sequential regression multiple imputation technique (also referred to as fully conditional specification) to impute HER2 status, ER status, PR status, and clinical/demographic variables with sporadic missingness (23). The idea behind this imputation technique is intuitive: to model each variable with missing observations conditional on all other variables (including those with missing values) and stochastically impute from these the conditional distributions. With a sporadic missingness pattern as is present in the SEER data, the imputation procedure cycles through these conditional models to produce a specified number of imputed datasets. Thus, the final imputed dataset would contain not only imputed HER2, ER, and PR statuses but also other variables that initially had missing information (e.g., stage, grade, node, etc.). Molecular subtype was not directly imputed, but rather derived for all cases based on observed and imputed HER2, ER, and PR status.

Supplementary Table S1 lists variables used for multiple imputation of missing HER2 status, along with the percentage of missing data for each variable. These variables were chosen based on demographic and clinical importance associated with HER2 status (2–4), and included year of diagnosis, age at diagnosis, poverty status (county level), urban indicator (county level), race, ethnicity, registry, reporting source, marital status, insurance, ER status, PR status, American Joint Committee on Cancer (AJCC) 7th stage, tumor size, nodal status, Bloom–Richardson grade, histology, survival time, vital status, and surgery. In addition, we used vital status and the Nelson–Aalen estimator of the cumulative baseline hazard in the imputation model for imputing missing HER2 status for the survival analyses (24). Poverty, urban, race, and surgery covariates had very little missing information (<1% missing). Therefore, we used a single imputation model controlling for age, Hispanic ethnicity, registry, and year of diagnosis to fill in missing information for these four covariates before performing multiple imputation on the remaining covariates in the breast cancer dataset. Because grade had the largest amount of missing information (i.e., 13.8% women with missing grade), we assessed the imputation procedure's sensitivity to inclusion or exclusion of this variable by repeating the imputation, excluding grade from the imputation model. We found no

Table 1. Demographic and clinical characteristics of molecular subtypes in women with invasive breast cancer (SEER-18 excluding Alaska, 2010–2013)

	All cases	Among cases with known subtype ($N = 179,098$)[a]				Among total cases[b]
		$HR^+/HER2^-$	Triple-negative	$HR^+/HER2^+$	$HR^-/HER2^+$	Unknown subtype
	$N = 196,094$	$N = 130,543$ (66.6%)	$N = 21,136$ (10.8%)	$N = 19,016$ (9.7%)	$N = 8,403$ (4.3%)	$N = 16,996$ (8.7%)
Demographic characteristics						
Age at diagnosis						
<50	44,648	26,819 (64.9%)	6,171 (14.9%)	5,970 (14.5%)	2,346 (5.7%)	3,342 (7.5%)
50–64	75,788	49,872 (71.4%)	8,400 (12.0%)	7,784 (11.2%)	3,748 (5.4%)	5,984 (7.9%)
65–74	42,713	30,857 (78.8%)	3,785 (9.7%)	3,154 (8.1%)	1,375 (3.5%)	3,542 (8.3%)
75+	32,945	22,995 (79.8%)	2,780 (9.6%)	2,108 (7.3%)	934 (3.2%)	4,128 (12.5%)
Race/ethnicity						
Non-Hispanic white	133,238	92,323 (75.6%)	12,710 (10.4%)	12,205 (10.0%)	4,932 (4.0%)	11,068 (8.3%)
Non-Hispanic black	21,739	12,035 (61.1%)	4,230 (21.5%)	2,283 (11.6%)	1,147 (5.8%)	2,044 (9.4%)
Non-Hispanic API	16,712	10,889 (71.3%)	1,430 (9.4%)	1,900 (12.4%)	1,062 (6.9%)	1,431 (8.6%)
Hispanic	22,373	14,088 (69.6%)	2,575 (12.7%)	2,417 (11.9%)	1,169 (5.8%)	2,124 (9.5%)
Poverty 2010 (county level)						
High SES	31,370	21,649 (75.2%)	2,923 (10.2%)	2,902 (10.1%)	1,321 (4.6%)	2,575 (8.2%)
Medium SES	136,612	91,229 (73.2%)	14,476 (11.6%)	13,213 (10.6%)	5,789 (4.6%)	11,905 (8.7%)
Low SES	28,065	17,638 (69.0%)	3,735 (14.6%)	2,899 (11.3%)	1,292 (5.1%)	2,501 (8.9%)
Urban 2010 (county level)						
Less urban (< 50%)	14,586	9,289 (70.8%)	1,790 (13.6%)	1,433 (10.9%)	612 (4.7%)	1,462 (10.0%)
More urban (≥ 50%)	181,461	121,227 (73.1%)	19,344 (11.7%)	17,581 (10.6%)	7,790 (4.7%)	15,519 (8.6%)
Insurance status						
Uninsured	4,062	2,255 (63.6%)	591 (16.7%)	470 (13.3%)	230 (6.5%)	516 (12.7%)
Any medicaid	23,167	13,919 (66.3%)	3,148 (15.0%)	2,585 (12.3%)	1,328 (6.3%)	2,187 (9.4%)
Insured	137,980	94,693 (74.0%)	14,315 (11.2%)	13,225 (10.3%)	5,658 (4.4%)	10,089 (7.3%)
Insured/no specifics	26,208	17,312 (73.7%)	2,736 (11.6%)	2,406 (10.2%)	1,046 (4.5%)	2,708 (10.3%)
Unknown	4,677	2,364 (74.3%)	346 (10.9%)	330 (10.4%)	141 (4.4%)	1,496 (32.0%)
Year of diagnosis						
2010	47,094	30,365 (72.1%)	5,233 (12.4%)	4,492 (10.7%)	2,000 (4.8%)	5,004 (10.6%)
2011	49,023	32,576 (73.0%)	5,398 (12.1%)	4,538 (10.2%)	2,089 (4.7%)	4,422 (9.0%)
2012	49,546	33,274 (72.9%)	5,322 (11.7%)	4,890 (10.7%)	2,150 (4.7%)	3,910 (7.9%)
2013	50,431	34,328 (73.4%	5,183 (11.1%)	5,096 (10.9%)	2,164 (4.6%)	3,660 (7.3%)
Marital status						
Single	29,188	18,327 (69.3%)	3,574 (13.5%)	3,205 (12.1%)	1,329 (5.0%)	2,753 (9.4%)
Married	105,955	71,630 (72.9%)	11,247 (11.5%)	10,724 (10.9%)	4,623 (4.7%)	7,731 (7.3%)
Separated	2,040	1,264 (68.4%)	264 (14.3%)	218 (11.8%)	103 (5.6%)	191 (9.4%)
Divorced	20,648	13,793 (72.4%)	2,367 (12.4%)	1,962 (10.3%)	921 (4.8%)	1,605 (7.8%)
Widowed	26,391	18,136 (77.3%)	2,481 (10.6%)	1,898 (8.1%)	940 (4.0%)	2,936 (11.1%)
Unknown	11,448	7,097 (73.2%)	1,160 (12.0%)	964 (9.9%)	471 (4.9%)	1,756 (15.3%)
Clinical characteristics						
AJCC 7th stage						
I	92,641	69,202 (80.4%)	7,076 (8.2%)	7,135 (8.3%)	2,632 (3.1%)	6,596 (7.1%)
II	62,905	40,280 (68.1%)	8,986 (15.2%)	6,960 (11.8%)	2,952 (5.0%)	3,727 (5.9%)
III	22,420	13,142 (62.4%)	3,274 (15.5%)	2,920 (13.9%)	1,740 (8.3%)	1,344 (6.0%)
IV	11,140	5,557 (60.3%)	1,311 (14.2%)	1,516 (16.4%)	837 (9.1%)	1,919 (17.2%)
Unknown	6,608	2,355 (67.1%)	462 (13.2%)	470 (13.4%)	223 (6.4%)	3,098 (46.9%)
Bloom–Richardson grade						
Low grade	44,032	38,652 (93.1%)	948 (2.3%)	1,551 (3.7%)	358 (0.9%)	2,523 (5.7%)
Medium grade	72,881	57,018 (82.7%)	3,349 (4.9%)	6,797 (9.9%)	1,743 (2.5%)	3,974 (5.5%)
High grade	52,033	22,156 (45.0%)	14,079 (28.6%)	8,137 (16.5%)	4,877 (9.9%)	2,784 (5.4%)
Unknown	27,148	12,717 (65.4%)	2,760 (14.2%)	2,531 (13.0%)	1,425 (7.3%)	7,715 (28.4%)
Tumor size						
<2.0 cm	109,855	80,779 (79.2%)	8,711 (8.5%)	9,058 (8.9%)	3,503 (3.4%)	7,804 (7.1%)
2.0–4.9 cm	61,224	37,756 (65.9%)	9,048 (15.8%)	7,286 (12.7%)	3,239 (5.6%)	3,895 (6.4%)
5.0+ cm	16,185	8,925 (60.6%)	2,643 (17.9%)	1,957 (13.3%)	1,211 (8.2%)	1,449 (9.0%)
Unknown	8,830	3,083 (61.9%)	734 (14.7%)	715 (14.4%)	450 (9.0%)	3,848 (43.6%)
Nodal status						
Positive	63,632	40,058 (67.1%)	7,713 (12.9%)	7,978 (13.4%)	3,976 (6.7%)	3,907 (6.14%)
Negative	127,240	88,843 (76.0%)	13,120 (11.2%)	10,713 (9.2%)	4,277 (3.7%)	10,287 (8.08%)
Unknown	5,222	1,642 (67.9%)	303 (12.5%)	325 (13.4%)	150 (6.2%)	2,802 (53.66%)

[a]Percent calculated among cases with a known breast cancer molecular subtype.
[b]Percent calculated among total cases.

major influence on the imputation results and therefore decided to keep grade in our final model.

We imputed missing HER2 status under the missing at random (MAR) assumption, as the standard sequential regression multiple imputation technique assumes the MAR mechanism. The MAR assumption is not testable; however, we evaluated known and unknown HER2 status by demographic and clinical variables used in the imputation model (Supplementary Table S2). Missing HER2 status appeared to be associated with several of these variables, thus providing plausibility for the MAR assumption

after conditioning on these associated variables with a missing pattern of HER2 status. We used *proc mi* with *fcs* in SAS 9.3 (SAS Institute) and generated 25 imputed datasets for analyses. We generated 25 imputations based on the rule that the number of imputations should be at least equal to the percentage of incomplete cases (23). Finally, we analyzed each imputed dataset separately, and then estimates were combined using Rubin's rules (25).

Survival analyses. We present estimates of 4-year breast cancer–specific survival according to molecular subtypes and clinical and demographic factors. Breast cancer–specific survival was calculated using the actuarial method with the SEER special COD variable (20). Deaths due to breast cancer were treated as the event and other causes of death as the censoring indicator. Survival times were censored at loss to follow-up, death from causes other than breast cancer, or December 31, 2014, whichever occurred first. Finally, a Cox proportional hazards model was used to evaluate the association between molecular subtypes and breast cancer–specific survival after controlling for demographic and clinical factors. Although a previous study showed survival curves by HR^+ or HR^- subtypes cross over when examining over a longer follow-up time (26), however, over the short follow-up period considered in this analysis, the proportionality assumption by molecular subtypes for the Cox model was tested [i.e., by looking at log-log (survival probability) over log (survival time) by molecular subtypes] and appeared to be valid.

Results

A total of 196,094 invasive breast cancer cases were diagnosed and reported to the SEER-18 excluding Alaska registries during 2010–2013. The proportions of women with each molecular subtype were 130,543 (66.6%) for $HR^+/HER2^-$, 21,136 (10.8%) for triple-negative (i.e., $HR^-/HER2^-$), 19,016 (9.7%) for $HR^+/HER2^+$, 8,403 (4.3%) for $HR^-/HER2^+$, and 16,996 (8.7%) unknown (Table 1). Subtype distributions varied by age, race, ethnicity, county-level poverty and urbanicity, insurance, marital status, stage, grade, tumor size, and nodal status. Compared with women with $HR^+/HER2^-$ subtype (the most common subtype), those diagnosed with the other three subtypes were somewhat more likely to be younger, belong to minority groups, living in counties with higher poverty levels, and had later stage, larger tumors, positive nodal status, and higher Bloom–Richardson grade.

Women with missing molecular subtypes tended to be older (12.5% missing among age 75+ vs. 7.5% age <50), minorities, poor, living in less urban areas, uninsured, and diagnosed with more advanced stage disease (17.2% stage IV vs. 7.1% stage I; Table 1). Supplementary Fig. S1 shows observed and imputed distributions for HER2 status by age at diagnosis based on one of the 25 imputed datasets (similar results were found when we repeated the analysis with a second imputed dataset). After imputation, the frequency distribution was shifted to the older age group from the younger age group (Supplementary Fig. S1A; Fig. 1B). This is expected because a larger percentage of older women were missing HER2 status than younger women. The percentages of $HER2^+$ and $HER2^-$ tumors were similar across ages before and after imputation (Supplementary Fig. S1C; Fig. 1D), even though older women had more missing information.

We present 4-year estimates of breast cancer–specific survival by molecular subtypes before and after imputation (Fig. 1). Before imputation, the survival curves were overestimated, whereas after imputation, the survival curves were shifted downward. This downward shift in survival after imputation is consistent with our finding that women with missing receptor status had worse prognostic features (Table 1; Supplementary Table S2; Supplementary Fig. S2). Based on

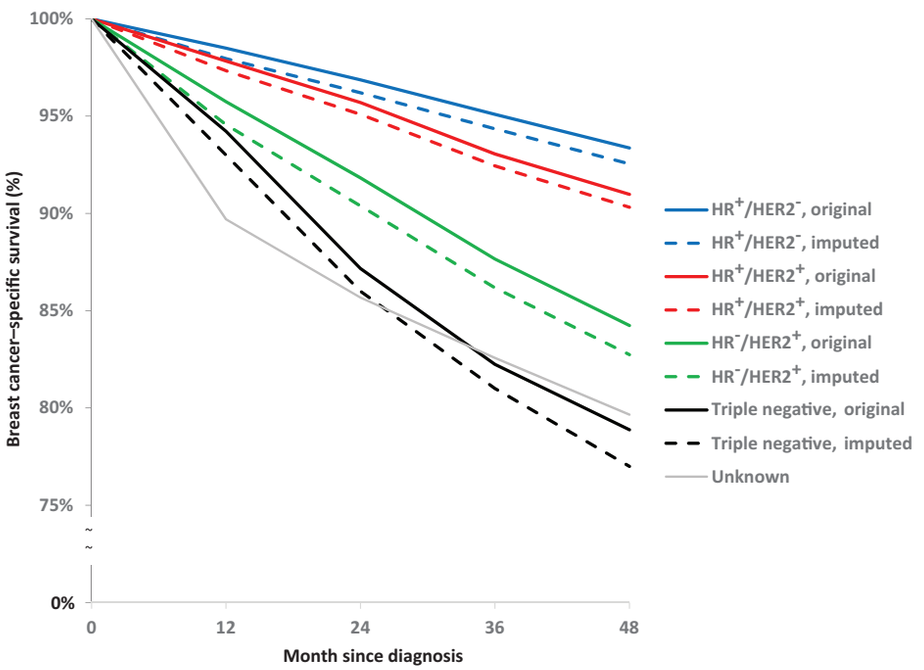

Figure 1.
Four-year breast cancer–specific survival by molecular subtypes before and after imputation, SEER-18 excluding Alaska. Breast cancer–specific survival curve using original data is shown with solid line for each of the five categories of molecular subtypes, whereas the corresponding survival curves using imputed data are shown with the dashed lines: $HR^+/HER2^-$ (blue solid and dashed lines), $HR^+/HER2^+$ (red solid and dashed lines), $HR^-/HER2^+$ (green solid and dashed lines), triple-negative (black solid and dashed lines), and unknown (gray solid line). Note that after imputation, there are no unknown subtypes reaming in the dataset.

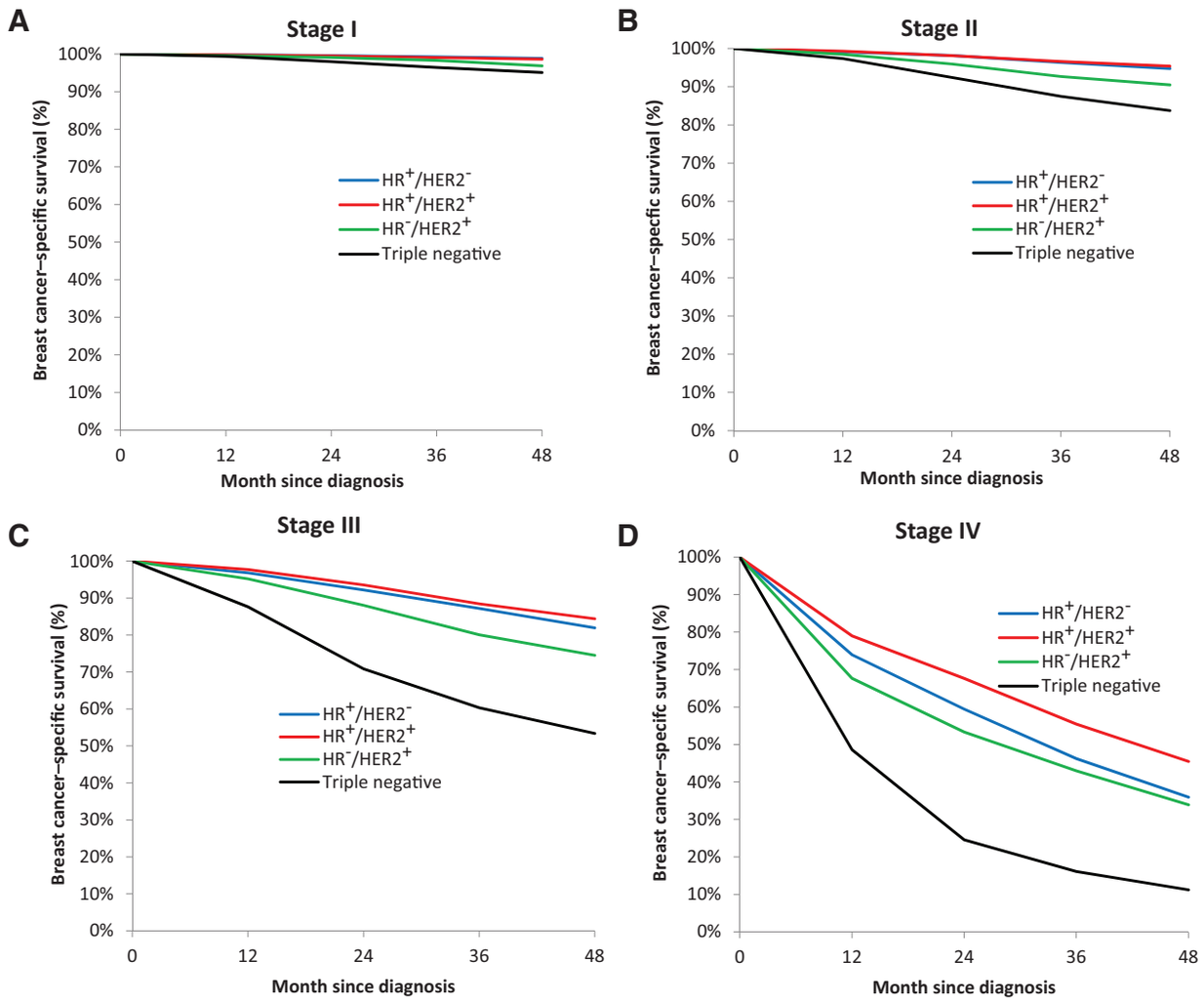

Figure 2.

Four-year breast cancer–specific survival by stage and molecular subtypes using imputed dataset, SEER-18 excluding Alaska. Breast cancer–specific survival curves using imputed data by stage at diagnosis and molecular subtypes are shown as follows: (**A**) among stage I disease, (**B**) among stage II disease, (**C**) among stage III disease, and (**D**) among stage IV disease; $HR^+/HER2^-$ (blue solid line), $HR^+/HER2^+$ (red solid line), $HR^-/HER2^+$ (green solid line), triple-negative (black solid line). Note that after imputation, there are no unknown subtypes or unknown stage groups reaming in the dataset.

imputed datasets, the best survival was observed among women with the $HR^+/HER2^-$ subtype (survival rate 92.5% at 4 years), followed by $HR^+/HER2^+$ (survival rate 90.3%), $HR^-/HER2^+$ (survival rate 82.7%), and finally worst survival for the triple-negative subtype (survival rate 77.0%).

Although molecular subtypes affected survival, stage at diagnosis appeared to be the most powerful factor (Fig. 2; Supplementary Fig. S2). For example, based on imputed datasets, the survival rate for the triple-negative subtype (known for poor prognosis) was 95.1% among stage I disease, yet dropped to 11.2% among those with stage IV disease (Fig. 2; Supplementary Table S3). Survival rates were similar between the $HR^+/HER2^-$ (the most common subtype known to have the best prognosis) and $HR^+/HER2^+$ subtypes among women with stage I and II disease but not stage IV disease, where survival rate was significantly higher for $HR^+/HER2^+$ than $HR^+/HER2^-$ subtypes (45.5% $HR^+/HER2^+$ vs. 35.9% $HR^+/HER2^-$). In Supplementary Table S3, we summarized survival estimates by molecular subtypes and other demographic and clinical characteristics.

After controlling for demographic and clinical factors, breast cancer–specific survival was significantly worse among the triple-negative subtype [HR, 2.5; 95% confidence interval (CI), 2.4–2.6], $HR^-/HER2^+$ (1.2; 95% CI, 1.1–1.3), but not for the $HR^+/HER2^+$ subtype (0.7; 95% CI, 0.7–0.8), compared with $HR^+/HER2^-$ as the referent subtype over this entire observation period (Table 2). Other factors associated with worse breast cancer–specific survival included advanced disease stage (32.7; 95% CI, 29.9–35.6), high grade (2.1; 95% CI, 1.9–2.2), non-Hispanic black race (1.1; 95% CI, 1.0–1.1), no surgery (2.8; 95% CI, 2.7–3.0), and poverty (low SES: 1.2; 95% CI, 1.1–1.3). Being married and having any type of insurance had a protective effect on survival.

Table 2. Cox model assessing factors associated with breast cancer–specific death (SEER-18 excluding Alaska, 2010–2013)[a]

Covariates	HR	95% CI	*P* value
Clinical characteristics			<0.0001
Breast subtypes			
$HR^+/HER2^-$	Ref.		
$HR^+/HER2^+$	0.77	(0.73, 0.83)	
$HR^-/HER2^+$	1.24	(1.16, 1.34)	
Triple-negative	2.52	(2.40, 2.65)	
AJCC 7th stage			<0.0001
I	Ref.		
II	3.62	(3.34, 3.91)	
III	11.67	(10.69, 12.73)	
IV	32.70	(29.98, 35.67)	
Bloom–Richardson grade			<0.0001
Low grade	Ref.		
Medium grade	1.30	(1.20, 1.41)	
High grade	2.12	(1.97, 2.29)	
Nodal status			>0.05
Negative	Ref.		
Positive	0.99	(0.94, 1.03)	
Surgery			<0.0001
Received surgery	Ref.		
Did not receive surgery	2.91	(2.78, 3.04)	
Demographic characteristics			
Age at diagnosis			<0.0001
<50	Ref.		
50–64	1.26	(1.20, 1.32)	
65–74	1.61	(1.52, 1.70)	
75+	3.07	(2.90, 3.24)	
Race/ethnicity			<0.0001
Non-Hispanic white	Ref.		
Non-Hispanic black	1.12	(1.06, 1.18)	
Non-Hispanic API	0.81	(0.75, 0.88)	
Hispanic	0.95	(0.89, 1.01)	
Poverty 2010 (county level)			<0.0001
High SES	Ref.		
Medium SES	1.15	(1.07, 1.22)	
Low SES	1.24	(1.14, 1.35)	
Urban 2010 (county level)			>0.05
Less urban (<50%)	Ref.		
More urban (≥50%)	0.99	(0.92, 1.06)	
Insurance status			<0.0001
Uninsured	Ref.		
Any Medicaid	0.86	(0.78, 0.94)	
Insured	0.63	(0.58, 0.69)	
Insured, NOS	0.72	(0.66, 0.80)	
Marital status			<0.0001
Single	Ref.		
Other[b]	0.90	(0.86, 0.94)	

[a]Cases were diagnosed from 2000 to 2013 and followed through December 31, 2014. We controlled for registry in the model.
[b]Other marital group contained married, partnered, and separated.

Discussion

This is the first study to assess breast cancer survival according to molecular subtypes and important clinical and demographic features using the largest population coverage to date at the national level in the modern treatment era and correcting for missing information with unknown receptor status to obtain a more accurate estimate of breast cancer prognosis in the population. We found that breast cancer cases with HR^+ subtypes are associated with the best prognosis, also shown in previous studies (17, 18, 27). By contrast, women with HR^- subtypes, especially those with triple-negative disease, suffer the worst prognosis, likely because of the lack of a receptor target (e.g., ER, PR, HER2) for therapy (18, 19, 28). In addition, stage is one of the most powerful factors determining survival outcomes. For example, among those with stage I disease, survival rate was greater than 95% regardless of subtypes after 4 years of follow-up. Among those with stage IV disease, women with the $HR^+/HER2^+$ subtype appeared to have better survival than women with the $HR^+/HER2^-$ subtype, even after controlling for other clinical and demographic factors. A similar pattern in terms of favorable survival with the $HR^+/HER2^+$ subtype was also shown using California Cancer Registry data (17, 18).

The triple-positive subtype has the advantage of expressing all therapeutic targets (ER, PR, HER2), which likely accounts for the superior survival we observed. Recently, a profusion of HER2-targeted therapies has entered clinical practice, including the combination of trastuzumab, pertuzumab, and chemotherapy which has yielded a median survival of nearly 5 years (56.5 months), substantially exceeding earlier median survival estimates in range of 2 years (8). Additional HER2-targeted options include trastuzumab emtansine and lapatinib (9, 10). These agents may be combined with endocrine therapy, chemotherapy, or each other, and are given sequentially in various permutations for many lines of treatment. Thus, treatment of the $HR^+/HER2^+$ subtype has advanced beyond what is available for other types of breast cancer. Our observation of a clearly superior survival with $HR^+/HER2^+$ disease in stage IV, and less so in earlier stages, might be due to the fact that stage I–III patients receive early HER2-targeted therapy, which may select for disease that is more resistant to HER2-directed therapies at the time of distant metastatic recurrence. Therefore, stage IV patients have the opportunity for maximal benefit from HER2-targeted therapy, because they have not already developed resistance to it.

This is the first time tumor receptor status was imputed for survival analyses using population-based registries. Use of multiple imputation to fill in missing receptor status is important. If we do not impute cases with missing receptor status, we overestimate survival because those with missing receptor status tend to have worse prognostic features. We used a rich set of covariates in the imputation model to make the MAR assumptions viable. Moreover, when we evaluated missing HER2 status by clinical, demographic, and breast cancer survival, missingness was explained by these factors, making MAR assumption more plausible. One limitation of our imputation model was that we did not have information on some other potential predictors of missing HER2 status such as breast cancer risk factors, treatment, and comorbid conditions. The SEER program recommends using imputed datasets when assessing breast cancer survival in the U.S. general population. These imputed datasets can be made available through SEER*Stat software to researchers in the future (29).

There are several strengths of our study. Our results are population-based and incorporate high-quality U.S. cancer registry data. SEER registries reliably capture breast cancer cases in their catchment areas, and they have complete follow-up information for greater than 95% of cases, so reporting of survival is reliable (30). Our results are more generalizable than those from single centers, and clinical trials are unlikely to include representative samples of older, sicker, and low-income patients (31, 32). Therefore, our study best reflects outcomes among unselected U.S. breast cancer cases experiencing typical patterns of care.

A few limitations of our study should be noted. Because we relied on data collected by cancer registries, we did not have detailed information on use of treatments including endocrine, HER2-directed therapy and chemotherapy, individual-level SES, breast cancer risk factors, or comorbid conditions. Controlling for these factors, if possible, would likely reduce confounding and improve our understanding of survival differences by subtype. The maximum follow-up time (59 months) was limited by the year in which relatively complete receptor data became available in SEER. In particular, over a decade, the use of Herceptin and other targeted therapy dramatically changed prognosis for breast cancer, especially for HER2-positive tumor. However, central cancer registries collected HER2 status beginning with cases diagnosed only recently (from 2010+), limiting our ability to assess long-term survival trend from breast cancer in the early part of Herceptin era.

In conclusion, we found that breast cancer prognosis in the U.S. population varies significantly based on molecular subtypes and associated clinical and demographic features. We also found that for *de novo* metastatic disease, women with the $HR^+/HER2^+$ subtype (once considered a poor prognostic feature) have better survival than those with the $HR^+/HER2^-$ subtype (often considered the best prognostic feature). This remarkable divergence of survival curves is likely attributable to major advances in HER2-targeted treatment (8–10, 33). As follow-up time increases, SEER data can be used to monitor and better understand the impact of how targeted therapies are contributing to reduce breast cancer mortality in the U.S. population.

Disclosure of Potential Conflicts of Interest

R. Andridge is a consultant/advisory board member for NCI. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: N. Howlader, K.A. Cronin
Development of methodology: N. Howlader, K.A. Cronin
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): N. Howlader
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): N. Howlader, R. Andridge
Writing, review, and/or revision of the manuscript: N. Howlader, K.A. Cronin, A.W. Kurian, R. Andridge
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): N. Howlader
Study supervision: N. Howlader

Acknowledgments

This study was supported by the Surveillance Research Program in the Division of Cancer Control and Population Sciences at the NCI.

Received July 11, 2017; revised August 30, 2017; accepted March 20, 2018; published first March 28, 2018.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
2. Kohler BA, Sherman RL, Howlader N, Jemal A, Ryerson AB, Henry KA, et al. Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. J Nat Cancer Inst 2015;107:djv048–djv.
3. Howlader N, Altekruse SF, Li CI, Chen VW, Clarke CA, Ries LAG, et al. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 Status. J Nat Cancer Inst 2014;106:dju055–dju.
4. Clarke CA, Keegan TH, Yang J, Press DJ, Kurian AW, Patel AH, et al. Age-specific incidence of breast cancer subtypes: understanding the black-white crossover. J Natl Cancer Inst 2012;104:1094–101.
5. Howlader N, Chen VW, Ries LA, Loch MM, Lee R, DeSantis C, et al. Overview of breast cancer collaborative stage data items–their definitions, quality, usage, and clinical implications: a review of SEER data for 2004-2010. Cancer 2014;120:3771–80.
6. Anderson WF, Rosenberg PS, Katki HA. Tracking and evaluating molecular tumor markers with cancer registry data: HER2 and breast cancer. J Nat Cancer Inst 2014;106:dju093–dju.
7. Keegan TH, Derouen MC, Press DJ, Kurian AW, Clarke CA. Occurrence of breast cancer subtypes in adolescent and young adult women. Breast Cancer Res 2012;14:R55.
8. Swain SM, Baselga J, Kim S-B, Ro J, Semiglazov V, Campone M, et al. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. N Engl J Med 2015;372:724–34.
9. Geyer CE, Forster J, Lindquist D, Chan S, Romieu CG, Pienkowski T, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med 2006;355:2733–43.
10. Verma S, Miles D, Gianni L, Krop IE, Welslau M, Baselga J, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. N Engl J Med 2012;367:1783–91.
11. Haque R, Ahmed SA, Inzhakova G, Shi J, Avila C, Polikoff J, et al. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. Cancer Epidemiol Biomark Prev 2012;21:1848–55.
12. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. Clin Med Res 2009;7:4–13.
13. Carey LA, Perou CM, Livasy CA, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 2006;295:2492–502.
14. Bauer KR, Brown M, Cress RD, Parise CA, Caggiano V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California Cancer Registry. Cancer 2007;109:1721–8.
15. Parise CA, Bauer KR, Brown MM, Caggiano V. Breast cancer subtypes as defined by the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) among women with invasive breast cancer in California, 1999–2004. Breast J 2009;15:593–602.
16. Bauer K, Parise C, Caggiano V. Use of ER/PR/HER2 subtypes in conjunction with the 2007 St Gallen Consensus Statement for early breast cancer. BMC Cancer 2010;10:228.
17. Tao L, Gomez SL, Keegan THM, Kurian AW, Clarke CA. Breast cancer mortality in African-American and non-Hispanic white women by molecular subtype and stage at diagnosis: a population-based study. Cancer Epidemiol Biomark Prev 2015;24:1039–45.
18. Tao L, Chu L, Wang LI, Moy L, Brammer M, Song C, et al. Occurrence and outcome of de novo metastatic breast cancer by subtype in a large, diverse population. Cancer Causes Control 2016;27:1127–38.
19. O'Brien KM, Cole SR, Tse C-K, Perou CM, Carey LA, Foulkes WD, et al. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. Clin Cancer Res 2010;16:6100–10.
20. Howlader N, Ries LAG, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. Improved estimates of cancer-specific survival rates from population-based data. J Nat Cancer Inst 2010;102:1584–98.
21. Howlader N, Noone AM, Yu M, Cronin KA. Use of imputed population-based cancer registry data as a method of accounting for missing information: application to estrogen receptor status for breast cancer. Am J Epidemiol 2012;176:347–56.
22. Andridge R, Noone A-M, Howlader N. Imputing estrogen receptor (ER) status in a population-based cancer registry: a sensitivity analysis. Stat Med 2017;36:1014–28.
23. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30:377–99.

24. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009;28:1982–98.
25. Little RJA RD. Statistical analysis with missing data. New York, NY: John Wiley & Sons, Inc; 2002.
26. Anderson WF, Chen BE, Jatoi I, Rosenberg PS. Effects of estrogen receptor expression and histopathology on annual hazard rates of death from breast cancer. Breast Cancer Res Treat 2006;100:121–6.
27. Li X, Yang J, Peng L, Sahin AA, Huo L, Ward KC, et al. Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. Breast Cancer Res Treat 2017;161:279–87.
28. Lund MJ, Trivers KF, Porter PL, Coates RJ, Leyland-Jones B, Brawley OW, et al. Race and triple negative threats to breast cancer survival: a population-based study in Atlanta, GA. Breast Cancer Res Treat 2009;113:357–70.
29. Surveillance Research Program DoCCaPS, National Cancer Institute. SEER*Stat Software. Bethesda, MD: National Cancer Institute; 2010.
30. Johnson CJ WH, Yin D, Niu X. The impact of patient follow-up on population-based survival rates. J Registry Manag 2010;37.
31. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: Race-, sex-, and age-based disparities. JAMA 2004;291:2720–6.
32. Unger JM, Hershman DL, Albain KS, Moinpour CM, Petersen JA, Burg K, et al. Patient income level and cancer clinical trial participation. J Clin Oncol 2013;31:536–42.
33. Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, et al. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. Lancet 2017;389:1195–205.