RESEARCH ARTICLE

Does Breast Cancer Drive the Building of Survival Probability Models among States? An Assessment of Goodness of Fit for Patient Data from SEER Registries

Hafiz Khan[1]*, Anshul Saxena[2], Abhilash Perisetti[3], Aamrin Rafiq[4], Kemesha Gabbidon[3], Sarah Mende[1], Maria Lyuksyutova[3], Kandi Quesada[1], Summre Blakely[1], Tiffany Torres[1], Mahlet Afesse[1]

Abstract

Background: Breast cancer is a worldwide public health concern and is the most prevalent type of cancer in women in the United States. This study concerned the best fit of statistical probability models on the basis of survival times for nine state cancer registries: California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington. **Materials and Methods:** A probability random sampling method was applied to select and extract records of 2,000 breast cancer patients from the Surveillance Epidemiology and End Results (SEER) database for each of the nine state cancer registries used in this study. EasyFit software was utilized to identify the best probability models by using goodness of fit tests, and to estimate parameters for various statistical probability distributions that fit survival data. **Results:** Statistical analysis for the summary of statistics is reported for each of the states for the years 1973 to 2012. Kolmogorov-Smirnov, Anderson-Darling, and Chi-squared goodness of fit test values were used for survival data, the highest values of goodness of fit statistics being considered indicative of the best fit survival model for each state. **Conclusions:** It was found that California, Connecticut, Georgia, Iowa, New Mexico, and Washington followed the Burr probability distribution, while the Dagum probability distribution gave the best fit for Michigan and Utah, and Hawaii followed the Gamma probability distribution. These findings highlight differences between states through selected sociodemographic variables and also demonstrate probability modeling differences in breast cancer survival times. The results of this study can be used to guide healthcare providers and researchers for further investigations into social and environmental factors in order to reduce the occurrence of and mortality due to breast cancer.

Keywords: Breast cancer data- survival- probability models- goodness of fit tests

Asian Pac J Cancer Prev, **17 (12)**, 5287-5294

Introduction

Cancer has adversely affected the lives of many individuals, resulting in high mortality and morbidity rates. Currently, cancer is the second leading cause of death in the United States, preceded only by heart disease. However, it is expected to become the top leading cause of death within the next few years (Siegel et al., 2015). Despite declines in the incidence of prostate and lung cancer, breast cancer rates have remained relatively stable. Today there are more than three million women living in the US with a history of invasive breast cancer (Siegel et al., 2013). Moreover, it is estimated that one in eight women living in the US will develop breast cancer in her lifetime (DeSantis et al., 2014).

In most instances the risk of developing cancer increases with age. Out of all women who have been diagnosed with breast cancer in the US, 40% of breast cancer cases occurred among women over age 65 and another 20% among women under age 50 (Siegel et al., 2013). It is estimated that 1 in 53 women ages 0-49 will be diagnosed with breast cancer, as well as 1 in 44 women ages 50-59, 1 in 29 women ages 60 to 69, and 1 in 15 women over 70 (Iqbal et al., 2015; Kwan et al., 2014; Siegel et al, 2015). Comparatively, younger women are more likely to be diagnosed with triple negative cancers (Clarke et al., 2012).

Population-based data on cancer incidence has been collected by the Surveillance, Epidemiology, and End Results since 1973 and the Centers for Disease Control and Prevention's National Program for Cancer Registry since 1995 (Siegel et al., 2015). SEER, however, is the only source providing long term delay-adjusted and population-based incidence data (Siegel et al., 2015). Data for the nine selected states in this study were provided by SEER: California, Connecticut, Georgia, Hawaii, Iowa,

*[1]Department of Public Health, [3]School of Medicine, Texas Tech University Health Sciences Center, Lubbock, Texas 79430, [2]Department of Health Promotion and Disease Prevention, Florida International University, Miami, FL 33199, [4]Department of Computer Science, Texas Tech University, Lubbock, Texas 70409, USA. *For Correspondence: hafiz.khan@ttuhsc.edu*

Michigan, New Mexico, Utah and Washington (Siegel et al., 2015). Using a joint-point regression model based on cancer incidence between 1997 and 2011, it was estimated that 231,840 women would be diagnosed with breast cancer by 2015 and that 40,290 deaths due to breast cancer would occur in 2015 (Siegel et al., 2015). The incidence of breast cancer from 2007 to 2011 for each state included in this paper are as follows: (a) Connecticut had 136.6 cases per 100,000 with a death rate of 20.8, (b) California had 122.4 cases per 100,000 with a death rate of 21.5, (c) Georgia had 123.8 cases per 100,000 with a death rate of 23.1, (d) Iowa had 124.8 cases per 100,000 with a death rate of 20.7, (e) Michigan had 120.7 cases per 100,000 with a death rate of 23.5, (f) Hawaii had 126.0 cases per 100,000 with a death rate of 26.8, (g) New Mexico had 110.0 cases per 100,000 with a death rate of 26.8, (h) Washington had 132.5 cases per 100,000 with a death rate of 21.1, and (i) Utah had 112.0 cases per 100,000 with a death rate of 20.8 (Siegel et al., 2015).

There have been significant declines in breast cancer-related death rates across the US. From 1997 to 2011, these declines in cancer death rates ranging from 25 - 30% occurred in Maryland, New Jersey, Massachusetts, New York, and Delaware. During the same years, another 20,000 deaths were averted in California due to a 25% drop in cancer incidence (Siegel et al., 2015). Southern states appeared to have the slowest declines and western states showed the lowest death rates (Ning et al., 2015; Siegel et al., 2015). These differences could reflect disparities in resources and prevention. In 2011, cancer incidence ranged from 125.6 cases per 100,000 in Utah to 200.9 cases per 100,000 in Kentucky (Siegel et al., 2015). Model-based methodology was used to predict breast cancer incidence and death for 2015. The estimated female breast cancer cases in 2015 for each of the nine states were as follows: (a) Hawaii at 1,140 cases, (b) New Mexico at 1,320 cases, (c) Utah at 1,490 cases, (d) Iowa at 2,390 cases, (e) Connecticut at 3,190 cases, (f) Washington at 5,480 cases, (g) Georgia at 7,170 cases, (h) Michigan at 7,780 cases, and (i) California at 25,270 cases. Additionally, the estimated female breast cancer deaths per state were as follows: (a) Hawaii at 130 deaths (b) New Mexico at 270 deaths, (c) Utah at 270 deaths, (d) Iowa at 390 deaths, (e) Connecticut at 460 deaths, (f) Washington at 830 deaths, (g) Georgia at 1,240 deaths, (h) Michigan at 1,410 deaths, and (i) California at 4,320 deaths. Estimated new cancer cases for women in the US in 2015 were 810,170 cases, with an estimated 29% attributed to breast cancer. The estimated cancer deaths for women in 2015 were 277,280, with an estimated 15% attributed to breast cancer. These declines in breast cancer incidence and mortality in the US are a result of prevention efforts and advancement in medical practices (Siegel et al., 2015).

Breast cancer incidence and mortality also varies between different ethnicities. From 2007 to 2011, the incidence of breast cancer was 127.6 cases per 100,000 among White non-Hispanics, 123.0 cases per 100,000 for Black non-Hispanics, 91.7 cases per 100,000 among American Indians and Alaska Natives, 91.6 cases per 100,000 among Hispanics, and 86.0 cases per 100,000 among Asians and Pacific Islanders. Additionally, the death rates among ethnic groups were 11.3 deaths per 100,000 among Asians and Pacific Islanders, 14.5 deaths per 100,000 among Hispanics, 15.2 deaths per 100,000 among American Indians and Alaska Natives, 22.2 deaths per 100,000 among White non-Hispanics, and 31.4 deaths per 100,000 among Black non-Hispanics. The five-year survival for White women had increased by 16% from 76% in 1975 to 92% in 2010 and the five-year survival of Black women increased by 18% from 62% in 1975 to 80% in 2010 (Siegel et al., 2015). However, another study showed that the frequency of breast cancer decreased among White non-Hispanic women but increased among Blacks, Asians, White Hispanics, and American Indians and Alaska Natives between January of 2000 and December of 2006 (Ooi et al., 2011). This could be due to numerous racial disparities. For instance, Blacks are more likely to receive poor care, face higher levels of poverty, receive a later stage diagnosis, and encounter lower stage-specific survival (Siegel et al., 2015). Moreover, the highest risk of death in Blacks is associated among those with co-morbidities. Co-morbidities, as well as disparities between racial and ethnic groups, are a predictors of survival (Ning et al., 2015). Black women are more likely to have higher co-morbidity scores compared to other racial and ethnic groups, which leads to a high risk of mortality associated with co-morbidity (Ning et al., 2015).

Compared to White non-Hispanics, other racial and ethnic groups have been shown to have greater odds of dying from breast cancer. In one study explaining the role of race and ethnicity in stage 1 breast cancer diagnosis and treatment, there were some differences identified. The majority of breast cancer diagnosis (71%) was seen in White non-Hispanics (Iqbal et al., 2015). After White-non-Hispanics, minority racial and ethnic groups have the greatest odds of dying from cancer (Ooi et al., 2011). Japanese women were more likely to be diagnosed at stage 1 breast cancer compared to Black women who were diagnosed at later stages (Iqbal et al., 2015). Additionally, among Asians, the Japanese had lower odds of breast cancer while Koreans had higher odds (Iqbal et al., 2015). Black women had the highest proportion of triple-negative breast cancer and were more likely to die within seven years. Black women had the highest odds of being diagnosed with estrogen receptor negative/progesterone receptor negative (ER-/PR-) cancers.

Blacks and White Hispanics had greater odds of receiving inappropriate surgical and radiation breast cancer treatment (Ooi et al., 2011). Mexicans, and South and Central Americans had a greater likelihood of being diagnosed with stage III or IV breast cancer and of receiving inappropriate treatment while Puerto Ricans were more likely to be diagnosed with ER-/PR- cancers and an increased risk of mortality (Ooi et al., 2011). Of all racial and ethnic groups, Samoans had greater odds of receiving inappropriate treatment and were more likely to be diagnosed at stage IV (Ooi et al., 2011). Overall, Blacks, Hispanics, Hawaiians, and American Indians presented with more advanced stages of breast cancer, ER-/PR-, and mortality from breast cancer (Ooi et al., 2011). Black women in particular, were adversely affected by lower utilization of screening and preventative services, poor

clinical care after subsequent breast cancer diagnosis, poor long-term follow-up care after an abnormal mammogram, and clinical management (Ooi et al., 2011 and Siegel et al., 2015). Alaskan Natives and American Indians were more likely to be diagnosed with more advanced ER-/PR- cancers (Siegel et al., 2015).

In a comparative study of 7,375 black women 65 years and older diagnosed with breast cancer between 1991 to 2005, and 7,375 matched white control patients, 12% of Blacks compared to 5% of Whites did not receive any care for their breast cancer (Silber et al., 2013). Also, the mean age of diagnosis was later for Blacks compared to Whites (Silber et al., 2013). This can be attributed to Blacks presenting with decreased health status at diagnosis, more advanced stages of the disease, worse biological features of cancer, and co-morbid conditions (Silber et al., 2013).

A study by Kwan et al. (2014) highlights some of the health disparities other ethnicities experience that could contribute to breast cancer mortality among these groups. This study investigated body size and survival categorized by race/ethnicity in 11,351 cancer patients diagnosed between 1993 and 2007. Study findings indicated that obese women were most likely to be African Americans, Hispanics, older, current and past smokers, less educated, US born, and living in lower socioeconomic status (SES) neighborhoods (Kwan et al., 2014). Hispanics, Whites, Blacks, and Asians tended to live with higher proportions of the population having less than a high school education (Ooi et al., 2011). These racial and ethnic groups had 1.5 to 2.4 higher odds of stage IV tumors and the same odds of being diagnosed with ER-/PR- cancers. Additionally, obese women were more likely to be diagnosed with advanced stage, poorly differentiated, and larger tumors (Kwan et al., 2014). They were also least likely to receive hormonal therapy and breast cancer surgery (Kwan et al., 2014). Interestingly, women with the highest and lowest BMI had a 1.4 greater risk of death by breast cancer (Kwan et al., 2014). Khan et al. (2014a, b, c, d) has conducted extensive work with SEER breast cancer data. These studies examine statistical analyses and survival probability among breast cancer patients between various states and ethnicities. The readers are referred to these works.

It is essential that trends in breast cancer patients' statistics need to be studied in order to reduce the prevalence of this disease in the US. The goals of this study are: (i) to acquire knowledge of descriptive statistics for some selected demographic and socioeconomic characteristics surrounding breast cancer incidence, and (ii) to identify probability-survival models by using goodness of fit tests for breast cancer patients.

Materials and Methods

Several skewed statistical probability models have been used for inferential statistics. This study presents the use of advanced statistical probability models, including Burr, Wakeby, Gen. Gamma, Gamma, Pearson 6, and Gen. Pareto. Various goodness-of-fit statistics such as Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared were used to determine the best probability distribution that fits the data.

Survival times from 2,000 breast cancer patients from each of the nine states (California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington) were used in the statistical data analysis. SPSS software was (version 21, IBM Inc., Chicago, IL, US, 2015) used to calculate the descriptive statistics for the selected socio-demographic variables and the EasyFit software (2015) was used for the assessment of goodness of fit testing and to identify the best fit survival models.

Data Source and Selection of Cases

The data for this study were obtained from the National Cancer Institute's SEER cancer registry program. The SEER program collects data through eighteen cancer registries in the US that are located in San Francisco-Oakland SMSA (1973), Connecticut (1973), Metropolitan Detroit (1973), Hawaii (1973), Iowa (1973), New Mexico (1973), Seattle (Puget Sound) (1974), Utah (1973), Metropolitan Atlanta (1975), Alaska (1992), San Jose-Monterey (1992), Los Angeles (1992), Rural Georgia (1992), Greater California (excluding San Francisco (SF), Los Angeles (LA) & San Jose (SJ))(2000), Kentucky (2000), Louisiana (2000), New Jersey (2000) and Greater Georgia (excluding Atlanta (AT) and Rural Georgia (RG)) (2000). Data files for breast cancer cases were downloaded from the SEER website. For this analysis, male breast cancer cases were excluded and we selected only female patients diagnosed with primary invasive breast cancer from years 1973 to 2012. A breakdown of total breast cancer cases for each registry is shown in Table 1. Breast cancer by state was our primary exposure of interest, and we selected individuals residing in the US according to the demographic information provided in the SEER (2010) database.

The specific inclusion criteria were as follows: (1) females; (2) Black, White, Other, and Hispanic race/ethnicity; (3) age of diagnosis between 18 and 90 years old; (4) diagnosis between 1973 and 2012 with breast cancer as the primary cancer diagnosis; (5) AJCC stages I to IV; (6) known tumor size; and (7) the degree of axillary lymph node involvement (LN), estrogen receptor (ER) and progesterone receptor (PR) statuses. Women who were diagnosed with breast cancer at death or by autopsy and those with other first primary cancers, in situ disease and no record of surgery type or radiation therapy were excluded from this analysis. Patients diagnosed with incomplete information were also not included.

Results

The total number of breast cancer cases reported between 1973 and 2012 was 1,385,980 (Table 1). The largest sample was drawn from the Greater California (excluding SF, LA and SJ) region with 199,716 (14.4%) cases and the lowest from Alaska with 1,352 (0.1%) cases. Among regions where data was available since 1973, Metropolitan Detroit had the greatest number of reported breast cancer cases with 123,936 (8.94%) females and Hawaii reported 31,415 (2.3%) breast cancer cases, which was lowest among this group. For the purpose of this study,

Table 1. Total Female Breast Cancer Cases (1973-2012)

Cancer Registry	Frequency	Percent (%)
San Francisco-Oakland SMSA (1973)	122,693	8.9
Connecticut (1973)	119,258	8.6
Metropolitan Detroit (1973)	123,936	8.9
Hawaii (1973)	31,415	2.3
Iowa (1973)	89,708	6.5
New Mexico (1973)	39,159	2.8
Seattle (Puget Sound) (1974)	113,129	8.2
Utah (1973)	37,653	2.7
Metropolitan Atlanta (1975)	63,555	4.6
Alaska*	1,352	0.1
San Jose-Monterey*	37,880	2.7
Los Angeles*	138,535	10.0
Rural Georgia*	2,081	0.2
Greater California (excluding SF, Los Angeles & SJ)**	199,716	14.4
Kentucky**	46,703	3.4
Louisiana**	46,514	3.4
New Jersey**	113,531	8.2
Greater Georgia (excluding AT and RG)**	59,162	4.3
Total	1,385,980	100%

(Year in parentheses refers to first diagnosis year data reported to SEER.); *Note,The incidence/yr1992_2012.sj_la_rg_ak directory files contain cases for Alaska, San Jose-Monterey, Los Angeles and Rural Georgia registries beginning in 1992; Cases have been collected by SEER for these registries prior to 1992 but have been excluded from the SEER Research Data file; **Note, The incidence/yr2000_2012.ca_ky_lo_nj_ga directory files contain cases for Greater California, Kentucky, Louisiana, New Jersey and Greater Georgia registries beginning in 2000; For the year 2005, only January through June diagnoses are included for Louisiana; The July through December incidence cases can be found in the yr2005.lo_2nd_ half directory

we excluded registries that had data only from the year 1990 onwards. The nine registries with data available from 1973 to 2012 included in the analysis were California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington.

A summary of descriptive statistics are shown in Table 2. The mean age of diagnosis across all nine states ranged from late 50s to mid-60s. States with the earliest ages of diagnosis included Georgia at a mean age of 58.0 years (SD = 14.0) and Hawaii with a mean age of 59.7 years (SD = 13.6), while Iowa reported the latest mean age of diagnosis at age 64.1 years (SD = 14.5). The state reporting the greatest variance in age of diagnosis was California with a mean of 61.4 years (SD = 25.2). Across the nine states, age of diagnosis spanned from ages 16 (Utah) to age 104 (Iowa). Other mean age of diagnosis were Connecticut at 62.5 years (SD = 14.3), Washington at 61.5 years (SD = 14.4), New Mexico at 61.3 years (SD = 14.1), Michigan at 60.98 years (SD = 14.2), and Utah at 60.8 years (SD = 14.5).

Patients' survival time was measured in months, post diagnosis. At approximately 8 years, New Mexico (105.0 months) and Georgia (105.7 months) reported the shortest survival time. While California reported the longest survival time (116.9 months) of almost 9.7 years, which was followed by Hawaii with a mean survival time of 114.1 months (9.5 years), and then by Iowa with 113.2 months (9.4 years), Washington 112.0 months (9.3 years), Connecticut 110.8 months (9.2 years), Michigan 108.2 months (9.0 years), and Utah 107.2 with months (8.93 years). The maximum patient survival time post diagnosis, was reported as 467 months (38.9 years) from Michigan.

Marital status was measured as single, married, separated, divorced, and widowed. California reported the highest number of single patients at 14.6%, followed by Georgia at 12.6%, then Connecticut at 11.7%, Michigan at 11.2%, Hawaii at 11%, New Mexico at 10%, Washington at 8.6%, Iowa at 7.2%, and Utah at 5.9%. The majority of patients across all of the nine states reported being married. Utah reported the highest percentage of married patients at 64.4%, followed by Hawaii at 59.6%, then by Iowa at 58.3%, Washington 58.0%, Georgia 54.8%, New Mexico 54.1%, California 53.1%, Connecticut 52.9%, and finally Michigan at 50.6%. The state with the highest number of patients categorized as separated was Connecticut at 4.3%, followed by California 1.3%, Georgia 1%, Hawaii 0.8%, Michigan 0.7%, New Mexico 0.5%, Iowa and Washington at 0.4%, and Utah 0.3%. Georgia reported the highest number of participants categorized as divorced at 11.8%, followed by Washington at 10.5%, then Michigan at 10.2% California at 9.7%, Hawaii 9.3%, Connecticut 8.6%, Utah at 8.3%, New Mexico 8%, and Iowa at 6.9%. Iowa reported the greatest percentage of widowed patients at 25.6%, followed by Michigan at 22.7%, then Utah at 18.9%, Connecticut at 18.8%, California at 18.5%, New Mexico at 18.2%, Washington at 18.1%, Georgia at 17.1%, and Hawaii at 16.4%.

Most states reported a sample of primarily White patients (>70.2%) with the exception of Hawaii reporting a sample of only 30.2% of White patients. Iowa reported

Table 2. Summary Statistics for Female Breast Cancer Patients of Nine States

Statistics	Registry								
Variables	California	Connecticut	Michigan	Hawaii	Iowa	New Mexico	Washington	Utah	Georgia
Age at Diagnosis (years)									
N (no. of patients)	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000
Mean	61.4	62.5	61.0	59.7	64.1	61.3	61.5	60.8	58.0
Std. Deviation	25.2	14.3	14.2	13.6	14.5	14.1	14.4	14.5	14.0
Minimum	24.0	25.0	22.0	24.0	23.0	17.0	26.0	16.0	23.0
Maximum	99.0	98.0	97.0	100.0	104.0	99.0	103.0	99.0	99.0
Survival time (months)									
N (no. of patients)	1,987	1,989	1,991	1,999	1,992	1,983	1,992	1,995	1,990
Mean	116.9	110.8	108.2	114.1	113.2	105.0	112.1	107.2	105.7
Std. Deviation	93.2	93.6	94.3	96.6	94.5	89.9	92.6	92.9	90.9
Minimum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum	462.0	459.0	467.0	462.0	458.0	464.0	449.0	459.0	443.0
Marital status									
Single	293 (14.6%)	233 (11.7%)	223 (11.2%)	219 (11%)	145 (7.2%)	199 (10%)	171 (8.6%)	119 (5.9%)	252 (12.6%)
Married	1,062 (53.1%)	1,059 (52.9%)	1,013 (50.6%)	1,192 (59.6%)	1,165 (58.3%)	1,082 (54.1%)	1,159 (58%)	1,288 (64.4%)	1,095 (54.8%)
Separated	25 (1.3%)	85 (4.3%)	14 (0.7%)	15 (0.8%)	8 (0.4%)	10 (0.5%)	9 (0.4%)	6 (0.3%)	19 (1%)
Divorced	193 (9.7%)	173 (8.6%)	203 (10.2%)	186 (9.3%)	138 (6.9%)	159 (8%)	210 (10.5%)	165 (8.3%)	235 (11.8%)
Widowed	370 (18.5%)	375 (18.8%)	454 (22.7%)	328 (16.4%)	512 (25.6%)	364 (18.2%)	361 (18.1%)	378 (18.9%)	342 (17.1%)
Unknown	57 (2.9%)	75 (3.8%)	93 (4.7%)	60 (3%)	32 (1.6%)	186 (9.3%)	90 (4.5%)	44 (2.2%)	57 (2.9%)
Race									
White	1,589 (79.5%)	1,876 (93.8%)	1,563 (78.1%)	604 (30.2%)	1,977 (98.9%)	1,894 (94.7%)	1,833 (91.6%)	1,945 (97.3%)	1,403 (70.2%)
Black	161 (8.1%)	105 (5.3%)	409 (20.5%)	17 (0.9%)	18 (0.9%)	17 (0.9%)	52 (2.6%)	9 (0.4%)	560 (28%)
Other	239 (12%)	9 (0.4%)	19 (1%)	1371 (68.6%)	2 (0.1%)	83 (4.2%)	114 (5.7%)	40 (2%)	34 (1.7%)
Unknown	11 (0.5%)	10 (0.5%)	9 (0.4%)	8 (0.4%)	3 (0.2%)	6 (0.3%)	1 (0.1%)	6 (0.3%)	3 (0.2%)
Ethnicity									
Non-Hispanic	1,860 (93%)	1,928 (96.4%)	1,975 (98.8%)	1,922 (96.1%)	1,992 (99.6%)	1,467 (73.4%)	1,973 (98.7%)	1,934 (96.7%)	1,946 (97.3%)
Hispanic	140 (7%)	72 (3.6%)	25 (1.3%)	78 (3.9%)	8 (0.4%)	533 (26.7%)	27 (1.4%)	66 (3.3%)	54 (2.7%)

Table 3. Goodness of Fit Analysis to Select Best Probability Survival Models

State	Probability Distribution	Kolmogorov Smirnov	Anderson Darling	Chi Squared	Parameter Estimates
California	Burr	0.046	51.7	116.9	k=2.3 α=1.3 β=195.0
	Dagum	0.028	43.6	41.5	k=0.3 α=2.9 β=189.9
	Wakeby	0.025	1.9	18.2	α=150.6 β=0.4 γ=5.1 δ=0.9 ζ=-1.1
Connecticut	Burr	0.043	43.8	102.1	k=2.7 α=1.2 β=221.8
	Dagum	0.028	35.6	16.3	k=0.3 α=2.9 β=191.3
	Wakeby	0.018	1.8	10.4	α=133.1 β=0.1 γ=0.1 δ=0.1 ζ=-2.1
Georgia	Burr	0.039	57.8	97.5	k=2.6 α=1.2 β=204.6
	Dagum	0.031	50.6	26.8	k=0.3 α=2.8 β=175.3
	Wakeby	0.016	1.1	8.8	α=126.1 β=0.2 γ=1.4 δ=0.1 ζ=-1.1
Hawaii	Gamma	0.029	92.4	37.2	α=1.1 β=110.0
	Gen Pareto	0.027	82.0	NA*	k=-0.1 s=126.2 m=0.8
	Wakeby	0.027	82.0	NA*	α=126.2 β=0.1 γ=0 δ=0 ζ=0.8
Iowa	Burr	0.033	58.5	60.4	k=2.9 α=1.3 β=239.5
	Gen Gamma	0.029	51.2	20.5	k=1.7 α=0.5 β=199.9
	Wakeby	0.013	0.8	7.5	α=139.2 β=0.2 γ=1.9 δ=0.9 ζ=-1.3
Michigan	Dagum	0.046	69.4	97.8	k=0.3 α=2.8 β=184.1
	Pearson 6	0.045	69.4	91.6	a1=1.3 α_2=3.9 β=288.3
	Burr	0.038	62.9	39.7	k=2.8 α=1.2 β=223.8
New Mexico	Burr	0.047	89.5	146.1	k=2.0 α=1.2 β=159.7
	Dagum	0.033	80.1	53.0	k=0.4 α=2.5 β=166.8
	Wakeby	0.030	3.6	34.6	α=123.3 β=0.3 γ=6.6 δ=0.9 ζ=-2.8
Utah	Dagum	0.037	47.3	66.2	k=0.3 α=3.1 β=186.7
	Burr	0.035	43.5	25.7	k=3.7 α=1.2 β=293.1
	Gen Gamma	0.030	42.2	22.1	k=1.6 α=0.5 β=188.2
Washington	Burr	0.041	56.5	85.3	k=3.0 α=1.3 β=243.4
	Dagum	0.035	49.3	26.6	k=0.3 α=3.2 β=195.9
	Gen Gamma	0.031	47.7	15.2	k=1.9 α=0.5 β=207.9

*Note, Probability distribution could not be determined by the software

the highest number of White patients at 98.9%, followed by Utah at 97.3%, New Mexico 94.7%, Connecticut 93.8%, Washington 91.6%, California 79.5%, Michigan 78.1%, and Georgia 70.2%. Georgia reported the highest number of Black patients at 28% followed by Michigan at 20.5%, California at 8.1%, Connecticut at 5.3%, Washington at 2.6%, Hawaii, Iowa and New Mexico reported 0.9%, and Utah at 0.4%. In Hawaii, 68.6% of patients were categorized as "other", followed by 12% in California, 5.7% in Washington, 4.2% in New Mexico, 2% in Utah, 1.7% in Georgia, 1% in Michigan, 0.4% in Connecticut, and 0.1% in Iowa.

Ethnicity was categorized as Hispanic or non-Hispanic. Most participants were categorized as non-Hispanic, with New Mexico reporting the lowest number of non-Hispanic patients at 73.4%. New Mexico was followed by California which reported 93% of non-Hispanic patients, then by Hawaii at 96.1%, Connecticut at 96.4%, Utah at 96.7%, Georgia at 97.3%, Washington at 98.7%, Michigan at 98.8%, and Iowa at 99.6%. New Mexico reported the highest number of Hispanic patients at 26.7% followed by California at 7%, Hawaii 3.9%, Connecticut at 3.6%, Utah 3.3%, Georgia at 2.7%, Washington 1.4%, Michigan 1.3%, and Iowa 0.4%.

The primary variable of interest was survival time (in months). The survival time was stratified by state registry and was imported into the EasyFit (2015) software to compare various statistical probability distributions that best fit the survival time data for each state. EasyFit (2015) builds the best fitting probability distribution based on the data and allows a large number of distributions to the data to fit. Variables that were included in our analyses were limited to those that are available in the SEER database. SPSS software, version 21 (2014) was used for all of the analyses.

Results from the goodness-of-fit analysis for the probability distributions from the software (Easyfit, 2015) are summarized in Table 3 as state or name of the registry, distribution name, Kolmogorov-Smirnov statistic, Anderson-Darling statistic, Chi-square value, and parameter values. The probability distributions identified by EasyFit were namely Burr, Dagum, Wakeby, Gamma, Gen. Pareto, Gen. Gamma, and Pearson 6. The three most commonly reported probability distributions were Burr, Dagum, and Wakeby. The data from eight out of nine registries was found to fit the Burr distribution

(California, Connecticut, Georgia, Iowa, Michigan, New Mexico, Utah, and Washington). New Mexico showed the highest Chi-square value among these groups for Burr distribution (146.1), the highest Kolmogorov-Smirnov value (0.04674), and the highest Anderson-Darling value (89.495). For the same distribution, Iowa had the lowest Chi-square value (60.38) and Kolmogorov-Smirnov value (0.03358), while Connecticut had the lowest Anderson Darling value (43.7). Data from seven registries fit the Dagum distribution (California, Connecticut, Georgia, Michigan, New Mexico, Utah, and Washington). For Dagum distribution, Michigan had the highest Chi-square value (97.8) and the highest Kolmogorov-Smirnov value (0.04643), whereas New Mexico had the highest Anderson-Darling value (80.1). The lowest values among the Dagum distribution included the Chi-squared value (16.33) and the Anderson-Darling value (35.636) for Connecticut and the Kolmogorov-Smirnov value for California (0.02799). Data from six registries fit Wakeby distribution (California, Connecticut, Georgia, Hawaii, Iowa, and New Mexico). New Mexico also had the highest Chi-square value (34.62), Kolmogorov-Smirnov value (0.03012) and Anderson-Darling value (3.6187) for Wakeby distribution, while Iowa had the lowest Chi-squared value (7.53), Kolmogorov-Smirnov value (0.01344), and Anderson-Darling value (0.83923). Utah also had the highest Chi-squared value (22.064) for Gen. Gamma distribution, and Washington had the lowest value (15.2). The Wakeby and Gen. Pareto distribution's Chi-square goodness of fit statistic for Hawaii was not available.

Discussion

Breast cancer remains a serious women's health and public health issue across the world and in the U.S. With millions of women predicted to be affected by this disease in the future, it is essential that more research over the epidemiology of this disease be addressed. Of particular interest are what groups of women (White, Black, Hispanic, Non-Hispanic, single, married, separated, divorced, widowed) are at higher risk of developing breast cancer and their respective survival times. Through our analysis of nine state registries with SEER (California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington) this study addressed descriptive statistics for each of the nine registries, as well as investigating the goodness-of-fit of possible probability models for breast cancer patient survival. Based on our findings, the Burr model fit six out of the nine registries being analyzed, with New Mexico showing the highest Chi-squared goodness-of-fit value at 146.09. Dagum probability model best fits for Michigan and Utah. Gamma probability model best fits for Hawaii. Additionally, California reported a mean survival of almost 9.74 years while New Mexico and Georgia reported a mean survival of almost 8 years. Clearly there are healthcare disparities that are influencing a gap of almost 2 years in breast cancer survival time post diagnosis.

It is imperative that future studies address possible policies, prevention programs, and healthcare disparities that could influence the rates of breast cancer incidence and mortality. Future studies could address the level of breast cancer education that breast cancer patients received prior to their diagnosis and during their diagnosis; this will aid in determining if the lack of breast cancer education pre-diagnosis is a leading factor in the lack of patients seeking out preventive screenings biannually. Further studies could also address how easily accessible women's health care centers are in different urban and rural areas. If women living in more rural areas compared to urban or suburban areas have later stage primary breast cancer diagnosis, or potentially shorter lifespans post diagnosis, this could indicate an issue in accessibility of care. These studies would address the availability to effective prevention programs. Socio-economic factors, as well as racial and ethnic discrimination, can be evaluated to gain more precise data on which groups of women are suffering from health care disparities leading to poor breast cancer screening rates and decreased survival time post-diagnosis. Understanding these contributing factors is necessary in order to create state-specific and population-specific programs and methods to decrease the prevalence of breast cancer as well as increase survival time post-diagnosis across all states in the United States.

SEER has collected cancer data for over thirty years from cancer registries throughout the United States; it is nationally recognized and considered a reliable source of information on incidence, mortality, and other related variables. Although the use of SEER lends this study strength, it is limited in the information it can provides. This studied is also limited by the length of time that is considered in this data set. It would be interesting future work for further analysis to be conducted for the states (California, Connecticut, Georgia, Iowa, New Mexico, and Washington) by splitting the data sets into multiple categories since there are numerous factors such as medical improvements, modern technology, and environmental changes which may influence the breast cancer survival probability. SEER lacks insight into many variables, such as social and economic factors that affect the survival time of breast cancer patients. Additionally, only data after the year 1992 was used for Alaska, San Jose-Monterey, Los Angeles, and Rural Georgia registries due to SEER excluding the cases prior to 1992 from the SEER Research Data file. Furthermore, only data after the year 2000 was used for Greater California, Kentucky, Louisiana, New Jersey, and Greater Georgia registries due to SEER excluding the cases prior to 2000 from the SEER Research Data file. For the year 2005, only January through June diagnoses were included for Louisiana in the SEER Research Data file.

The present study included some variables from SEER data, such as age at diagnosis, survival times, marital status, race, and ethnicity. For future studies socioeconomic factors may be considered to investigate disparities associated with breast cancer since they vary geographically according to different health policies.

Acknowledgements

The authors are grateful to the National Cancer

Institute in the United States for giving them access to the SEER's database. There are no conflicts of interest.

References

Co-Clarke CA, Keegan THM, Yang J, et al (2012). Age-specific incidence of breast cancer subtypes: understanding the black-white crossover. *J Natl Cancer Inst*, **104**, 1094-101.

DeSantis C, Ma J, Bryan L, Jemal A (2014). Breast cancer statistics, 2013. *CA Cancer J Clin*, **64,** 52–62.

EasyFit (2015). Distribution fitting made easy, mathwave-data analysis & simulation, version 5.5, retrieved on July 20, 2015 from http://www.mathwave.com/easyfit-distribution-fitting.html.

Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA (2015). Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. *JAMA*, **313**, 165.

Kwan ML, John EM, Caan BJ, et al (2014). Obesity and mortality after breast cancer by Race/Ethnicity: The California breast cancer survivorship consortium. *Am J Epidemiol*, **179**, 95–111.

Khan HMR, Ibrahimou B, Gabbidon K, et al (2014). a. statistical estimates from black non-Hispanic female breast cancer data. *Asian Pac J Cancer Prev*, **15**, 1-6.

Khan HMR, Saxena A, Gabbidon K, Ross E, Shrestha A (2014), b. Statistical applications for the prediction of white Hispanic breast cancer survival. *Asian Pac J Cancer Prev*, **15**, 5571–5.

Khan HMR, Saxena A, Gabbidon K, Stewart TSJ, Bhatt C (2014). c. Survival analysis for white non-Hispanic female breast cancer patients. *Asian Pac J Cancer Prev*, **15**, 4049–54.

Khan HMR, Saxena A, Vera V, et al (2014). d. black Hispanic and black non-Hispanic breast cancer survival data Analysis with half-normal model application. *Asian Pac J Cancer Prev*, **15**, 1–6.

Ning J, Peng S, Ueno N, et al (2015). Has racial difference in cause-specific death improved in older patients with late-stage breast cancer?. *Ann Oncol*, **26**, 2161-68.

Ooi SL, Martinez ME, Li CI (2011). Disparities in breast cancer characteristics and outcomes by race/ethnicity. *Breast Cancer Res Treat*, **127**, 729-38.

SEER (2010). Surveillance, epidemiology, and end results (SEER) program (www.seer.cancer.gov) research data (1973-2009), national cancer institute, DCCPS, surveillance research program. [Accessed July 10, 2013].

Siegel R, Desantis C, Virgo K, et al (2013). Cancer treatment and survivorship statistics , 2012. *CA Cancer J Clin*, **62**, 220-41.

Siegel R, Miller K, Jemal A (2015). Cancer statistics , 2015 . *CA Cancer J Clin*, **65**, 5-29.

Silber JH, Rosenbaum PR, Clark AS, et al (2013). Characteristics associated with differences in survival among black and white women with breast cancer. *JAMA*, **310**, 389.

SPSS (2015). Statistical package for the social sciences (SPSS), IBM Inc., Retrieved on February 15, 2015 from http://www-03.ibm.com/software/products/en/spss-statistics