RESEARCH ARTICLE

Statistical Applications for the Prediction of White Hispanic Breast Cancer Survival

Hafiz Mohammad Rafiqullah Khan[1]*, Anshul Saxena[2], Kemesha Gabbidon[2], Elizabeth Ross[3], Alice Shrestha[1]

Abstract

Background: The ability to predict the survival time of breast cancer patients is important because of the potential high morbidity and mortality associated with the disease. To develop a predictive inference for determining the survival of breast cancer patients, we applied a novel Bayesian method. In this paper, we propose the development of a databased statistical probability model and application of the Bayesian method to predict future survival times for White Hispanic female breast cancer patients, diagnosed in the US during 1973-2009. Materials and Methods: A stratified random sample of White Hispanic female patient survival data was selected from the Surveillance Epidemiology and End Results (SEER) database to derive statistical probability models. Four were considered to identify the best-fit model. We used three standard model-building criteria, which included Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Deviance Information Criteria (DIC) to measure the goodness of fit. Furthermore, the Bayesian method was used to derive future survival inferences for survival times. Results: The highest number of White Hispanic female breast cancer patients in this sample was from New Mexico and the lowest from Hawaii. The mean (SD) age at diagnosis (years) was 58.2 (14.2). The mean (SD) of survival time (months) for White Hispanic females was 72.7 (32.2). We found that the exponentiated Weibull model best fit the survival times compared to other widely known statistical probability models. The predictive inference for future survival times is presented using the Bayesian method. Conclusions: The findings are significant for treatment planning and health-care cost allocation. They should also contribute to further research on breast cancer survival issues.

Keywords: Breast cancer survival data - Bayesian inference - statistical models - survival inference

Asian Pac J Cancer Prev, **15 (14)**, 5571-5575

Introduction

Despite recent advances in the diagnosis and treatment of breast cancer, 458,400 lives were lost in 2008 to the illness globally (ACS, 2011; Bray et al., 2013). In the same year, approximately 1.4 million cases of breast cancer occurred worldwide (ACS, 2011). Furthermore, 50% of all breast cancer cases and 60% of all breast cancer associated deaths occurred in developing nations (ACS, 2011; Jemal et al., 2011). As of 2011, one in three women in the United States (U.S.) afflicted by cancer, specifically suffered from breast cancer (CDC, 2013). In addition, it was the second leading cause of cancer among women in the U.S., having an incidence rate of 123.1 cases per 100,000 (DeSantis et al., 2011a; NCI, 2012; CDC, 2013). The American Institute for Cancer Research (AICR) has predicted that about 226,000 cases of breast cancer will be diagnosed in the U.S. and approximately 18% of those cases will result in death (AICR, 2012). According to the American Cancer Society, there is a 16% prevalence of all types of breast cancer (ACS, 2013).

Breast cancer develops from the uncontrolled growth of altered breast tissue cells, developing into a tumor that is recognized as a lump or mass (NCI, 2012). Most breast cancers are carcinoma in situ (CIS) as they are confined only to the duct (Ductal Carcinoma in Situ, DCIS) or lobule (Lobular Carcinoma in Situ, LCIS). Major risk factors of breast cancer are smoking, excessive alcohol drinking, obesity, and a family history of breast cancer (Sexton et al., 2011; Zhao et al., 2013). Based on current breast cancer rates and screening capacity, the American Cancer Society recommends that asymptomatic women between the ages of 20 and 39 receive a clinical breast examination every three years, and women over age 40 are to receive an annual clinical breast examination and mammogram (Smith et al., 2010; DeSantis et al., 2011b).

Breast cancer has great variability among ethnic and racial groups accounting for differences in clinical manifestation, incidence, and disease prognosis (Sexton et al., 2011; NCI, 2012). Furthermore, health disparities are

[1]Department of Biostatistics, [2]Department of Health Promotion & Disease Prevention, Robert Stempel College of Public Health & Social Work, Florida International University, Miami, [3]Behavioral Science Research Corporation, Coral Gables, Florida, USA
**For correspondence: hmkhan@fiu.edu*

clearly identified by differences in socio-economic status, level of awareness, number of mammograms, and lack of access to health care. These factors are believed to affect one's likelihood of breast cancer diagnoses. Between the years 2004 and 2008, breast cancer incidence rates remained relatively stable among all racial and ethnic groups, and breast cancer death rates have decreased since the early 1990s among all ethnic groups except the American Indians and Alaska Natives (NCI, 2012). In the U.S., White women have the highest incidence rate of breast cancer at 124 cases per 100,000 (CDC, 2013). Breast cancer incidence is highest among White non-Hispanic women at 125.4 cases per 100,000, followed by African American at 116.1 cases per 100,000, Asian American and Pacific Islanders at 84.9 cases per 100,000, American Indian and Alaska Natives at 89.2 cases per 100,000, and Hispanics at 91.0 cases per 100,000 (DeSantis et al., 2011a; NCI, 2012). Though rates of breast cancer have declined overall, White non-Hispanic breast cancer rates increased for women ages 60 to 69 by 4.8% in 2007 (DeSantis et al., 2011b).

The American Cancer Society reports that disparities are evident among breast cancer death rates by state, socioeconomic status, race, and ethnicity (NCI, 2012). Based on the potential combined negative effects of these determinants it is important to understand their role in the breast cancer epidemic among American women. Currently the most predictive factor for breast cancer is age, as 78% of new cases and 87% of breast cancer deaths occurred in women over the age of 50 (DeSantis et al., 2011b). It however remains important to account for the disparities associated with race and ethnicity.

Despite the significant increase of Hispanics in the U.S., there is little known about the unique variations of breast cancers among this population. Breast cancer death rates among Hispanics are highest compared to all ethnicities (CDC, 2013). Furthermore, Hispanic women have less access to care and are typically unaware of their risk for developing breast cancer compared to women of other ethnic groups (NCI, 2012). In previous studies, Mexican-American women were considered a low risk group for developing breast cancer, however new data indicate higher incidence rates of breast cancer among this group. In addition, the Arizona Cancer Center collaborated with three Mexican universities on the Ella Binational Breast Cancer Study to collect information on breast cancer differences between Mexican native and Mexican-American women (NCI, 2012). Preliminary findings indicated that Mexican women who live in the U.S. have lifestyle and reproductive factors that increase their likelihood of developing breast cancer. These reproductive and lifestyle differences for Mexican-American women include the increased likelihood of beginning menstruation before age 12, alcohol consumption, obesity, and the use of hormone replacement therapy (NCI, 2012). Interestingly, this finding has been consistent with all other U.S. born Hispanic women. Comparatively, women born in Mexico have more children, breastfeed for a longer period, are more physically active, and consume more fiber, decreasing their risk of developing breast cancer (NCI, 2012).

There are many government agencies, hospitals, and other institutions that are collecting cancer survival data and storing them in computers for their future records. Statistical analyses can be applied to these existing data to make inferences. Statistical probability models are playing key roles in data analysis. Numerous statistical probability models are generated from data depending upon the discrete and continuous patterns.

Several statistical probability models have been applied to the health sciences field enabling researchers to better understand and make decisions using data. Healthcare researchers have applied these methods to various subjects, utilizing enormous amount of data brought about by the modern advancements in the biological sciences field. Although a large number of statistical models are used in data analysis, we have chosen to use the exponentiated exponential (EEM), exponentiated Weibull (EWM), beta generalized exponential (BGEM), and beta inverse Weibull model (BIWM) because of setting specific value of the parameters, several right skewed statistical probability models can easily be obtained.

The exponentiated exponential model (EEM) has been used in modeling data from biomedical sciences. The EEM has two parameters: scale and shape, where $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively (Khan et al., 2014a).

The exponentiated Weibull model (EWM) was applied to the examination of the breaking strength of materials. The probability model for the EWM is defined by three parameters, where $\alpha > 0$ and $\beta > 0$ are the shape parameters, and $\lambda > 0$ is the scale parameter (Khan et al., 2014a, 2014b).

The probability model of the beta generalized exponential model (BGEM) is defined by four parameters, where $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively. Introducing skewness and varying tail weight are functions of the two additional parameters, $a > 0$ and $b > 0$ (Barreto-Souza et al., 2010).

The probability model of the beta inverse Weibull model (BIWM) is defined by three parameters, where β is the shape parameter, and two additional parameters, $a > 0$ and $b > 0$ whose role is to introduce skewness and to vary tail weight (Khan et al., 2014c).

In order to explore the posterior probability for the parameters we used the Bayesian method. In this method, the likelihood is viewed as a function of parameter conditioned on a fixed observed data set. Prior data is investigated as a probability distribution and contains the parameter value (s) information. Multiplying the likelihood and 'prior' yields the joint distribution of the parameters that combine all the information about the parameters.

As a novel approach, the Bayesian method has much applicability in biomedical sciences. For more information about the Bayesian method and predictive inference have been discussed by a number of authors, Khan, 2012a-b; Khan, 2013a-b, among others.

The specific goals of the study are: *i)* to study some demographic and socio-economic variables of the selected sample; *ii)* to review the right skewed models EE, BGE, EW, and BIW; *iii)* to give a justification that the given

sample data follows a specific model by using model selection criterions for goodness of fit tests; *iv)* to obtain predictive inference for future survival time given the selected model.

This paper is organized as follows. Section 2 contains a detailed discussion of a real breast cancer survival data example and statistical methods. The measures of goodness of fit tests for the survival times of the race/ethnicities are presented in Section 3. Section 4 addresses the survival inference given the survival data points from the selected EWM. In Section 5, we discuss the results. Finally, Section 6 provides a discussion of overall findings.

Materials and Methods

The 657,712 breast cancer patients in the data were extracted from the Surveillance, Epidemiology and End Results website (SEER, 2012). We then applied stratified random sampling scheme to draw the sample from nine randomly selected states in order to represent White Hispanic Breast cancer cases. The SEER data consisted of 4,269 males and 653,443 females, males were excluded from the study because of the low probability of developing breast cancer. Furthermore, there were 22,639 White Hispanic and 531,562 White non-Hispanic women. We used a simple random sampling technique (SRS) method to select a sample of size 2,000 White Hispanic females from a total of 22,639 White Hispanic female cancer patients.

SPSS software (IBM, 2010) was used to generate descriptive statistics. Mathematica version 8.0 (Wolfram, 2012), computational software package was used to obtain predictive inference for future survival time. Finally, WinBugs an advanced software was used to verify the goodness of fit tests (Lunn et al., 2012).

Fitting of a Data-based Statistical Model

Akaike Information Criterion (AIC), Deviance Information Criterion (DIC), and Bayesian Information Criterion (BIC) are among the most popular advanced statistical methods used for measuring the goodness of fit of models. Goodness of fit tests is used to identify how well a statistical model fits the data. Of the three, the DIC is strongly preferred and is a Bayesian measure of fit that compares different models. DIC can have both negative and positive values. However, a model with a lower DIC value is considered better than others. As in the case of AIC, given any two estimated models, the model with lower value of BIC is preferred. To obtain AIC, BIC, and DIC values, one would consider the log-likelihood functions for the models. Four types of advanced models are used in breast cancer survival data as presented in the following Table 3.

Table 1. Frequency Distribution of Selected White Hispanic Breast Cancer Patients

States	White Hispanic Count	Percentage
Georgia	88	4.4
Hawaii	26	1.3
Iowa	29	1.4
Michigan	75	3.8
New Mexico	714	35.7
Utah	114	5.7
Washington	87	4.4
California	604	30.1
Connecticut	263	13.2
Total	2,000	100

Table 3 presents of AIC, BIC, and DIC values for the four models under study namely EEM, EWM, BGEM, and BIWM. In this table, the goodness of fit of survival times for White Hispanic female patients is tested. The fit of the model is determined by the values of criterion under study with the lowest values suggesting a better fit. According to the table above, the estimated values of both AIC and DIC are the lowest (19425.700 and 19423.700 respectively). In addition, in the estimated value of BIC, the value 19442.001 is very close to the lowest value of 19441.602. Since EW generates the smallest estimated values of all AIC, BIC, and DIC as compared to other models, it has the best fit for the survival times.

Survival inference

We used the Bayesian method to develop a predictive survival model for the survival times of the study sample, which is discussed in this section. As mentioned in section 3 the breast cancer survival data best fits the EW

Table 2. Statistics Results of Age at Diagnosis, Survival Time, and Marital Status at Diagnosis of Female White Hispanic Breast Cancer Patients

Characteristics	Categories	White Hispanic
Age at diagnosis (years)	Mean	58.17
	SD	14.18
	Median	57
	Range	17-100
	Quartile1	47
	Quartile2	57
	Quartile3	69
	Variance	201.03
Survival time (months)	Mean	72.7
	SD	32.17
	Median	74
	Range	31-142
	Quartile1	49
	Quartile2	74
	Quartile3	104
	Variance	1035.2
Marital status at diagnosis	Single	269
	Married	1054
	Separated	35
	Divorced	230
	Widowed	297
	Unknown	115

Table 3. Selection of the Best Model for White Hispanic Females on the Basis of AIC, BIC, and DIC Criterions

Model criterions	AIC	BIC	DIC
Exponentiated exponential	19430.4	19441.602	19430.426
Exponentiated Weibull	19425.7	19442.001	19423.7
Beta generalized exponential	19433.3	19455.703	19429.3
Beta inverse Weibull	19444.4	19465.7	19442.3

distribution based on the lowest value of model criterions.

Suppose data represent n White Hispanic female breast cancer patients' survival times that follow the EWM. The Bayesian posterior probability model can be defined by multiplying the likelihood function and prior for the parameters. Based on the n survival data points the likelihood function is the n times product of the fitted EW model. Furthermore, the Bayesian predictive model from the Weibull life model by means of a conjugate prior for the scale parameter and a uniform prior for the shape parameter is derived by Khan et al., 2011. Including Khan's et al. assumption about the prior knowledge for the parameters, the predictive summary results are obtained and reported in Table 4.

Results

A sample of 2,000 White Hispanic female breast cancer patients diagnosed during 1973 to 2009 was extracted from the SEER data. We used a stratified simple random sampling design to draw samples from randomly selected nine states. Varying statistical models were applied to identify the best-fit model for the survival data of the study sample. The majority of the sample consisted of patients from New Mexico (35.7%) followed by California (30.1%). Conversely, the least number of cases was found in Hawaii (1.3%) and Iowa (1.4%). The mean (SD), of age at diagnosis for the study sample was 58.17 (14.18) years. The minimum age at breast cancer diagnosis was 17 years and the maximum was 100 years. The mean (SD), of survival time was 72.70 (32.17) months and the survival time ranged from 31 to 142 months. The majority of the sample was married at the time of diagnosis.

The goodness of fit of the survival times for the study sample was tested. We analyzed the Exponential exponential model (EEM), Exponential Weibull model (EWM), Beta generalized exponential model (BGEM), and Beta inverse Weibull model (BIWM) by testing based on the three different criterions; AIC, BIC, and DIC.

Table 4. Predictive Inference Based on the EWM for White Hispanic Breast Cancer Patients Survival Data

	Summary	Statistics
	Mean	77.2893
	SE	0.765889
Raw moments	m_1	77.2843
	m_2	7146.04
	m_3	779741
	m_4	9.91838×10^7
Corrected moments	μ_1	77.2843
	μ_2	1173.17
	μ_3	46128.9
	μ_4	7.2053×10^6
Skewness & Kurtosis	β_1	1.31783
	β_2	5.23514
	γ_1	1.14797
	γ_2	2.23514
Survival intervals	90%	(25.7613, 127.3007)
	95%	(22.8667, 145.1649)
	98%	(17.0799, 170.4519)
	99%	(15.2961, 190.8772)

According to the obtained results, the Weibull distribution displayed the lowest AIC and DIC estimated values with BIC estimated value very close to the minimum value obtained. As we know that the better fit of the model is reflected by the lowest values, the Exponential Weibull model has the best fit and hence is the best model for the White Hispanic females.

According to the predictive results, the future survival time is higher for the study sample and is positively skewed. Predictive inference based on the EW model for the ethnic group under study is reported in Table 4 including the predictive mean, standard error (SE), raw moments, corrected moments, skewness and kurtosis, and survival intervals.

Discussion

Statistical modeling uses the application of statistical rules and restrictions to determine the model that best fits the data, this is a contrast to descriptive statistics that only allows basic interpretation of the data. In order to determine the best-fit model, tests measuring the goodness of fit are important. In this study, we used three model selection criterions, AIC, BIC, and DIC to develop a statistical probability model. Nine out of twelve states in the U.S. were selected and a stratified random sample of breast cancer patients was identified. In addition, we developed the fitted statistical survival model and derived the posterior distribution of the parameter by using the Bayesian method.

Several studies have used descriptive statistics to analyze survival data, however none used models to predict survival times. There is a great need for the use of predictive inferences to address future direction of disease. Based on the results presented in Table 4, the shape of the future survival models for the study sample is positively skewed. Study findings can provide practical assistance to healthcare researchers and medical providers for predicting a patient's possible future survival outcomes given the patient's past and current medical profile. Hence the findings, will effectively integrate the knowledge, discovery, and innovation contributing to an enhanced and improved rationale for the diagnosis and treatment of breast cancer patients throughout the nation, and potentially the world.

Acknowledgements

The authors would like to thank the editor and the referees for their valuable comments and suggestions.

References

American Cancer Society (2011). Global Cancer Facts & Figures, 2nd Edition. Retrieved from: http://www.cancer.org/research/cancerfactsfigures/globalcancerfactsfigures/global-facts-figures-2nd-ed.

American Cancer Society (2013). Cancer Facts and Figures. Retrieved from: http://www.cancer.org/research/cancerfactsstatistics/breast-cancer-facts-figures.

American Institute for Cancer Research (2012). Breast Cancer.

Retrieved from: http://www.aicr.org/learn-more-about-cancer/breast-cancer/.

Barreto-Souza W, Santos AHS, Cordeiro GM (2010). The beta generalized exponential distribution. *J Stat Comput Sim*, **80**, 159-72.

Bray F, Ren JS, Masuyer E, Ferlay J (2013). Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer*, **132**, 1133-45.

Centers for Disease Control and Prevention (2013). United States cancer statistics: 1999-2009. Retrieved from: www.cdc.gov/uscs.

DeSantis C, Siegel R, Bandi P, Jemal A (2011a). Breast cancer statistics, 2011, *CA*, **61**, 408-18.

DeSantis C, Howlader N, Cronin KA, Jemal A (2011b). Breast cancer incidence rates in US women are no longer declining. *Ca Epi Bio Prev*, **20**, 733-9.

IBM Corp (2010). IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY.

Jemal A, Bray F, Center MM, et al (2011). Global cancer statistics. *CA*, **61**, 69-90.

Khan HMR, Albatineh AN, Alshahrani S, Jenkins N, Ahmed NU (2011). Sensitivity analysis of predictive modeling for responses from the three-parameter Weibull model with a follow-up doubly censored sample of cancer patients. *Comput Stat Data An*, **55**, 3093-103.

Khan HMR (2012a). Estimating predictive inference for responses from the generalized Rayleigh model based on complete sample. *J Thai Statistician*, **10**, 53-68.

Khan HMR (2012b). Several priors based inference from the exponential model under censoring. *JP J Fund Appl Stat*, **2**, 1-13.

Khan HMR (2013a). Comparing relative efficiency from a right skewed model. *JP J Biostat*, **9**, 1-26.

Khan HMR (2013b). Inferential estimates from the one-parameter half-normal model. *J Thai Statistician*, **11**, 77-95.

Khan HMR, Saxena A, Rana S, Ahmed NU (2014a). Bayesian modeling for male breast cancer data. *Asian Pac J Cancer Prev*, **15**, 663-9.

Khan HMR, Saxena A, Kemesha G, Rana S, Ahmed NU (2014b). Model-based survival estimates of female breast cancer data. *Asian Pac J Cancer Prev*, **15**.

Khan HMR, Saxena A, Shrestha A (2014c). Posterior inference for white hispanic breast cancer survival data. *J Biomet Biostat*, **5**, 183.

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D (2012). A practical introduction to bayesian analysis, CRC Press / Chapman and Hall.

National Cancer Institute (2012). What you need to know about cancer: understanding cancer. retrieved from: http://www.cancer.gov/cancertopics/types/breast.

Sexton KR, Franzini L, Day RS, et al (2011). A review of body size and breast cancer risk in Hispanic and African American women. *Cancer*, **117**, 5271-81.

Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW (2010). Cancer screening in the United States, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. *CA*, **60**, 99-119.

Surveillance, Epidemiology and End Results (2012). Cancer of the breast - seer stat fact sheets. retrieved from http://seer.cancer.gov/statfacts/html/breast.html.

Wolfram Research (2012). The Mathematica Archive: Mathematica 8.0. Wolfram Research Inc, Illinois.

Zhao G, Li C, Okoro CA, et al (2013). Trends in modifiable lifestyle-related risk factors following diagnosis in breast cancer survivors. *J Cancer Survivorship*, 1-7.