https://doi.org/10.33472/AFJBS.6.6.2024.6979-6986

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Predicting Healthcare Utilization among Breast Cancer Patients with Several Regression Modeling Approach

[1]Dr. Mrinal Deka, [2]Dr. Parbin Sultana

[1]Faculty of Science, Statistics, Assam down town University, Guwahati - 781026
[2]Professor, School of Technology and Management, University of Science and Technology Meghalaya

Article Info

Volume 6, Issue 6, July 2024

Received: 23 March 2024

Accepted: 20 June 2024

Published: 09 July 2024

doi: 10.33472/AFJBS.6.6.2024.6979-6986

ABSTRACT:

The research has successfully picked out various risk factors like Socio-Economic, Demographic and Clinical for breast cancer. It divulges that patient age, tumour size, node size, and blood sugar levels have a negative impact on cancer patient survival times. Specifically, as these factors increase, survival times decrease. Additionally, the study addresses the crucial issue of selecting an appropriate survival model. It determines that the Weibull survival model is the most suitable, as it shows lower AIC values compared to other models for breast cancer. The findings suggest that the estimated survival times from this model are reliable, enabling predictions of breast cancer patient's survival times based on available data.

Key Words: Breast cancer, Survival analysis, Risk factors, Weibull survival model, Prediction

1. INTRODUCTION

Cancer is widely conceded as a significant public health bother, causing considerable suffering and reducing the lifespan of affected individuals. Despite its profound societal impact, this aspect is often underestimated (Chu et al., 2008). However, it would be imprudent to solely analyse socio-economic factors without considering demographic and clinical factors.

Cancer circumscribes a group of diseases distinguished by the unusual growth of tissues, forming tumours (Baghestani et al., 2015). The nature and outcomes of cancer differ depending on its position within the body (Ali et al., 2011). Yet, there has been scanted research into understanding this heterogeneity, which could greatly inform treatment procedures and medical direction.

Breast cancer highlights as a major cause of female mortality globally (Uysal, 2017). A meta-analysis of 119 studies involving over 12 million women disclosed more than 2,60,000 outlined cases of breast cancer (Yeole et al., 2003).

Understanding the circumstances and treatment of different types of cancer require an evaluation of various risk factors, including socio-economic, demographic and clinical factors. Age, for instance, is a globally acknowledged risk factor for cancer incidence, with most cancers becoming more prevalent with increasing age (White et al., 2014). Additionally, research indicates discrepancies in cancer mortality rates between genders, with men more susceptible to certain types of cancers (Kim et al., 2018). Economic factor also plays a significant role, as the price of cancer treatment often leads to financial distress, especially for economically indigent individuals, influencing their access to early perception and treatment (Nair et al., 2014).

Early perception is paramount in intercepting breast cancer development, yet lack of consciousness about risk factors and a little access to healthcare among economically indigent populations often result in detained diagnosis (Caplan, 2014; Walter et al., 2015). Adequate sensation and financial care can facilitate quick diagnosis and treatment initiation, thereby better survival rates (Zhai et al., 2019).

Investigating disease incidence data using suitable statistical tools yield precise insights even in the face of unpredictability (Mohamad et al., 2007). Consequently, this research employs different survival models to analyse the effect of risk factors on breast cancer patient survival, focusing on recognize the most suitable model for the available data and forecast survival times effectively.

While supporting literature furnishes meaningful insights, there persists a gap in studies that concurrently consider several risk factors, involving socio-economic, demographic, and clinical factors, especially in regions like South Assam. In addition to this, predictive models for cancer patient survivability comprising these factors are short falling for such high-risk areas. Therefore, the current research aims to address these crannies by analysing the interaction of numerous risk factors and demographic information in forecasting cancer patient consequences in South Assam.

2. METHOD AND MATERIAL

The study depends on secondary data origin from the Cachar Cancer Hospital and Research Centre, situated in Silchar, Assam, spanning from 2014 to 2019. The hospital's first principle is to convey excellent treatment to its patients while supporting moderate service costs. It is noted that 80 percent of the patients are engaged in daily wage jobs, with 50 percent earning nominal wages *(https://cacharcancerhospital.org)*. The majority of hospitalized patients usually hail from different districts across Assam and other northeastern states such as Tripura, Manipur, Meghalaya, and Mizoram (Ngaihte et al., 2019 and Mathur et al., 2020).

Total of 377 patients with breast cancer were selected observing the prevalent occurrence of this type of cancer in the region.

The present study designs to explore the correlation between socio-economic, demographic, and medical factors and the survival rates of breast cancer patients. Econo-demographic factors such as age, gender, marital status, religion, and consumption habits are considered alongside

medical factors like size of tumour, node, level of blood sugar and treatment procedures. These factors are treated as independent variables (risk factors), while the survival time of cancer patients considers as the dependent variable in constructing the model.

Before formally consolidating these factors into survival models, it is necessary to assess whether few independent variables play redundant roles, probably affecting the reliability of the statistical models. Therefore, correlations are computed among both the dependent and independent variables, as well as among the independent variables themselves. Point Bi-serial correlation and Karl Pearson correlation are computed for categorical and continuous variables respectively. In addition to this, Collinearity Diagnostics are conducted to measure the degree of multicollinearity within the model.

Parametric survival models incorporating the Exponential, Weibull and Gaussian models are applied to analyse the effects of various factors on the survival times of breast cancer patients. The selection of the suitable survival model for the dataset is determined using the Akaike Information Criterion (AIC), where a lower AIC value indicates a better-fitted model for the data.

Furthermore, Bayesian regression modeling is used to authenticate the estimates, acquired from the selected survival model. Bayesian linear regression, originate from the Bayesian approach, indicates uncertainty as probability, conflict with the frequentist approach. This approach integrates prior information about parameters before perceiving the data, using prior distributions. The posterior distribution, defining the updated optimisms after observing the data, is acquired using the likelihood function. Markov Chain Monte Carlo (MCMC) method is generally utilized to approximate the posterior distribution. Mathematically, the posterior distribution can be determined as the distribution of unknown parameters given the observed data.

$$Posterior \propto Likelihood \times Prior \ldots (1)$$

We can also write equation (1) as

$$P(\theta|D) \propto P(D|\theta) \times P(\theta) \ldots (2)$$

Thus, based on Bayes theorem, we can write the following equation

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{\sum P(D|\theta) \times P(\theta)} \ldots (3)$$

It is more convenient if we write the equation (3) as follows

$$P(Model|New_{Data}) = \frac{P(New_{Data}|Model) \times P(Model)}{P(New_{Data})} \ldots (4)$$

In this study, a non-informative prior was used due to the absence of sufficient knowledge about the parameters, thereby considering them neutral.

$$\beta_j \sim N(\mu_j, \sigma_j^2) \ldots (5)$$

In our analysis, we applied the 'brms' package, version 4.0.2, in R to derive posterior distributions for Bayesian Regression Models using Stan.

Our primary objective was to forecast survival times using appropriate survival models. To accomplish this, we adopted a holdout sample technique, segregating 30 samples as holdout samples for breast cancer, while the remaining samples considered as training data. Subsequently, we computed estimated survival times from the holdout samples, distinctively for cancer site and contrasted them with actual survival times. In addition, we calculated the Mean Square Error (MSE) to compute the average squared difference between the estimated and actual survival times, providing insight into the predictive performance of our models.

3. RESULT

Initially, correlations were computed to get redundancy among the dependent and independent variables, resulting in the construction of (Table 1). While prior literature specifies the significance of all variables as potential risk factors, this research will compass solely on Age, Size of Tumour, Node and Blood Sugar Level, as they generate significant correlations with the survival time of patients for breast cancer (as shown in Table 1). Other factors will be excluded from the model as their correlations with the dependent variable are not statistically significant, recommending potential redundancy in the model (Uyanik et al., 2013).

Table 1: Correlation between dependent and independent variable of breast cancer

Type of cancer	Factor	Correlation Coefficient	*p* value
Breast cancer	Surv_Time and Age	-.072	.013
	Surv_Time and Tumour Size	-.011	.006
	Surv_Time and Node Size	-.663	.001
	Surv_Time and Blood Sugar Level	-.203	.001
	Surv_Time and Marital Status	-.825	.098
	Surv_Time and Religion	.656	.069
	Surv_Time and Consumption Habit	.726	.078
	Surv_Time and Treatment	-.652	.089
	Surv_Time and Economic Status	.796	.099

Again, we assess the correlation among the chosen independent variables to identify and eliminate redundancy within the regression model, resulting in the construction of the subsequent table.

Table 2: Correlations between the selected independent of breast cancer site

Type of cancer	Independent Variable	Correlation Coefficient	*p* value
Breast cancer	Age and Tumour_size	-.057	.266
	Age and Blood Sugar Level	.047	.168
	Age and Node size	-.051	.145
	Tumour_size and Blood Sugar Level	.040	.436
	Tumour_size and Node_size	-.082	.114
	Node_size and Blood Sugar Level	-.747	.123

The correlations between the picked independent variables for breast cancer as shown in Table 2, indicate non-significance, with p-values exceeding 0.05. Therefore, we consider Age, Size of Tumor, Node and level of blood sugar as the independent variables. Collinearity diagnostics between these variables were also computed, obtaining from the following table.

Table 3: Collinearity diagnostics for the chosen independent variables of breast cancer

Type of Cancer	Independent Variable	VIF value	Condition Index (CI)
Breast Cancer	Age	1.356	7.562
	Tumour size	1.585	8.962
	Node size	2.370	6.320
	Blood sugar level	1.120	4.852

The table (Table 3) represents collinearity diagnostics for the selective independent variables of breast cancer, indicating that multicollinearity is effectively overseed with Variance Inflation Factor (VIF) values below 5 and Condition Indices below 15. On the other hand, parametric survival models such as the 'Exponential,' 'Weibull' and 'Gaussian' models were applied to investigate the effect of factors on the survival times of breast cancer patients. To determine the most appropriate survival model for the dataset, the Akaike Information Criterion (AIC) was used, where a lower AIC value represents a better-fitted model for the selective data (Kleinbaum and Klein, 2012). The next table is constructed to aid in this gauging.

Table 4: Calculated AIC values for the three parametric survival models

Type of Cancer	Survival Model	AIC Value
Breast Cancer	Weibull Model	2604.464
	Exponential Model	3236.481
	Gaussian Model	2652.534

Table 4 designates that the Weibull survival model is the most suitable for this study, as it shows lower AIC values compared to other survival models of breast cancer. Consequently, we solely compass on the regression estimates derived from the Weibull survival model.

Table 5: Outcomes derived from the Weibull survival model for various cancer sites

Type of Cancer	Factor	Estimate	Standard Error	p value	95 % Credible Interval	
					LCL	UCL
Breast Cancer	Intercept	5.3083	0.1055	.013	5.1015	5.5151
	Age	-0.0038	0.0011	.006	-0.0060	-0.0015
	Tumour_size	-0.0187	0.0156	.001	-0.0494	0.0119
	Node_size	-0.0214	0.0114	.001	-0.0254	0.0345
	Blood Sugar Level	-0.0137	0.0006	.023	-0.0150	-0.0124

Based on the findings outlined in Table 5, we have pinpointed independent variables that wield a significant impact on the dependent variable, with p-values below 0.05. Notably, the coefficients associated with the independent variables—patient age, size of tumour, node and level of sugar are consistently negative. This urges a negative influence on the survival time of breast cancer patients. Substantially, as age of the patient, size of tumour, node and level of blood sugar increase, the survival times of cancer patients tend to decrease.

Moreover, to sustain the reliability of our estimates derived from the Weibull survival model, we have applied a Bayesian regression model. This additional analysis focuses to support the agility of our findings, thereby furnishing further credibility to our results. The outcomes of this Bayesian regression model are condensed in the following table.

Table 6: Estimates derived from both the Weibull survival model and Bayesian regression model

Type of Cancer	Factor	Estimate (Obtained from Weibull model)	Estimate (Obtained from Bayesian model)
Breast Cancer	Intercept	5.3083	5.2000
	Age	-0.0038	-0.0031
	Tumour_size	-0.0187	-0.0200
	Node_size	-0.0214	-0.0219
	Blood Sugar Level	-0.0137	-0.0128

After acquiring parameter estimates from both models, as mentioned in Table 6, we observe that the estimates are almost identical. Successively, our confidence in the unspecified parameter estimates is strengthen by the new knowledge acquired from observed data. Our key objective is to predict survival times using the Weibull survival model. To accomplish this, we apply the holdout sample technique, dedicating 30 samples for breast cancer while utilizing the remainder as training samples. We then compute estimated survival times for the holdout samples, comparing them with the actual survival times. Additionally, we calculate the Mean Square Error (MSE) to obtain the average squared difference between the estimated and actual survival times. The subsequent figures provide the comparison of estimated and actual survival times for Breast cancer.

Fig. 1: Comparison of estimated and actual survival times for Breast cancer

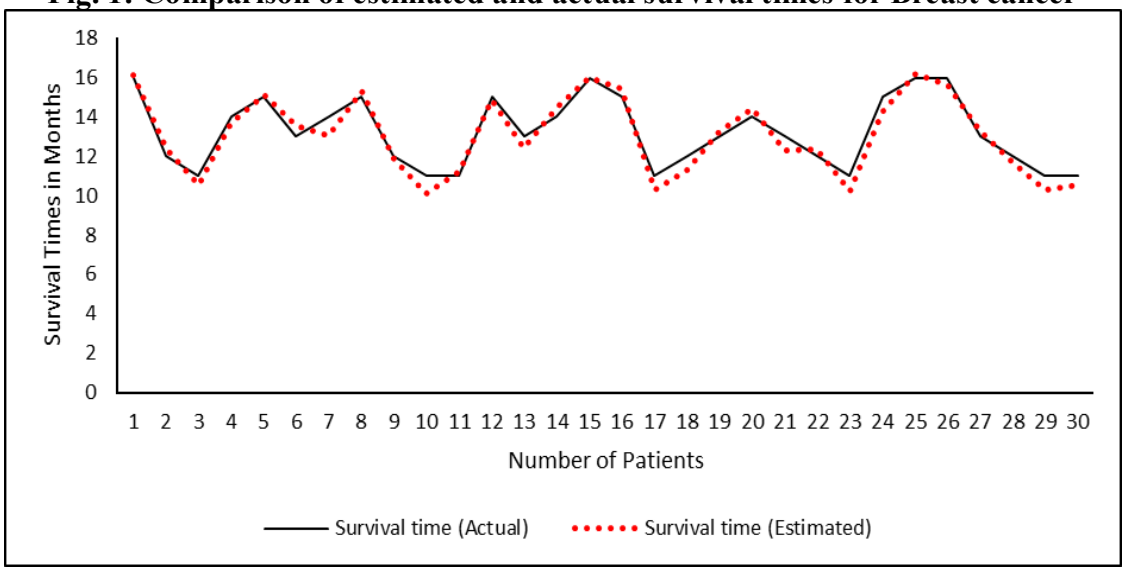

Figure 1 explains minimal variance between the actual and estimated survival times for breast cancer (Mean Squared Error = 0.2431). Consequently, we can affirm that the estimated survival times derived from the model demonstrate reliability. Thus, the model holds promise for predicting survival times for breast cancer patients based on the provided dataset.

4. CONCLUSION

The research has effectively identified associations between different risk factors, including socio-economic, demographic and clinical factors with breast cancer. Our findings reveal that age of patient, size of tumour, node and levels of blood sugar have a detrimental impact on

cancer patient survival times. Specifically, as these factors increase, survival times decrease. Moreover, the study addresses the important question of choosing an appropriate survival model for our dataset. We determine that the Weibull survival model is the most suitable model, as it shows a lower AIC value compared to other models. In addition to this, we obtain that the estimated survival times derived from this model are dependable and can be used for predicting survival times among cancer patients based on the available data.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

Supplementary materials

There is no Supplementary material for this article.

5. REFERENCE

1. Ali I, W. W. (2011). Cancer Scenario in India with Future Perspectives. *Cancer Therapy*, 50-70.
2. Baghestani A, M. S. (2015). Survival Analysis of Patients with Breast Cancer using Weibull Parametric Model. *Asian Pacific Journal of Cancer Prevention*, 8567-8571.
3. Caplan L. (2014). Delay in Breast Cancer: Implications for Stage at Diagnosis and Survival. *Frontiers in Public Health*, 1-5.
4. Chu P, W. J. (2008). Estimation of Life Expectancy and the Expected Years of Life Lost in Patients with Major Cancers: Extrapolation of Survival Curves under High-Censored Rates. *Value in Health*, 1102-1109.
5. Kim I. H, H. L. (2018). Sex Differences in Cancer: Epidemiology, Genetics and Therapy. *Biomolecules & Therapeutics*, 335-342.
6. Kleinbaum D.G, K. M. (2012). *Survival Analysis, A Self-Learning Text* (3rd ed.). Springer.
7. Mathur P, S. K. (2020). Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. *American Society of Clinical Oncology*, 1063-1075.
8. Mohamad P, E. H. (2007). Comparing Cox Regression and Parametric Models for Survival of Patients with Gastric Carcinoma. *Asian Journal of Cancer Prevention*, 412-416.
9. Nair S. K, R. S. (2014). Cost of Treatment for Cancer : Experiences of Patients in Public Hospitals in India. *Asian Pacific Journal of Cancer Prevention*, 5049-5054.
10. Ngaihte P, E. Z. (2019). Cancer in the NorthEast India: Where We are and What Needs to Be Done? *Indian Journal of Public Health*, 251-253.
11. Uyanik K. G, G. N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 234-240.
12. Uysal E. (2017). Top 100 Cited Classic Articles in Breast Cancer Research. *European Journal of Breast Health*, 129-137.
13. Walter M. F, G. R. (2015). Symptoms and other Factors Associated with Time to Diagnosis and Stage of Lung Cancer: A Prospective Cohort Study. *British Journal of Cancer*, 6-13.
14. White C. M, D. M. (2014). Age and Cancer Risk: A Potentially Modifiable Relationship. *American Journal of Preventive Medicine*, 7-15.

15. Yeole B. B, K. A. (2003). An Epidemiological Assessment of Increasing Incidence and Trends in Breast Cancer in Mumbai and Other Sites in India, During the Last Two Decades. *Asian Pacific Journal of Cancer Prevention*, 51-56.
16. Zhai Z, F. Z. (2019). Effects of Marital Status on Breast Cancer Survival by Age, Race and Hormone Receptor Status: A Population-Based Study. *Cancer Medicine*, 4906–4917.