International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Survival Analysis of Cancer Patients Using Weibull Parametric Model

Wilson Kiprotich Chepkech[1], Joel Cheruiyot Chelule[2], Ayubu Anapapa[3], Herbert Imboga[4]

[1, 2, 4] Jomo Kenyatta University of Agriculture and Technology, Department of Statistics and Actuarial Science, Nairobi, Kenya

[3]University of Eldoret, Department of Mathematics and Computer Science, Eldoret, Kenya

Abstract: ***Cancer incidence and mortality are rapidly growing worldwide. The reasons are complex but reflect both aging and growth of the population, as well as changes in the prevalence and distribution of the main risk factors for cancer, several of which are associated with socioeconomic development. This study was analysed using Weibull distribution model on the survival of cancer patients. This study is inspired by Siegel R. L, et al (2017) .The main aim of this study was to analyse the survival of cancer patients using Weibull distribution model. The general objective of this study was to analyse the survival of cancer patients using Weibull parametric model. Analysis was performed using R and other mathematical modules were discussed. From the analysed data, age of diagnosis, Histologic grade, Cancer type, BMI, stages of infections and Gender were significant factors in the survival of Cancer patients. Family history and treatment administered did not show much significant effect. In future studies, using other parametric distributions such as gamma Weibull distributions which leads to cover different types of hazard functions were suggested.***

Keywords: Weibull distribution model, Survival Time, Cancer disease

1. Introduction

Cancer is a generic term for a large group of diseases characterized by the growth and spread of abnormal cells beyond their usual boundaries that can then invade adjoining parts of the body and/or spread to other organs (WHO 2017). Cancer arises from the transformation of normal cells into tumour cells in a multistage process that generally progresses from a pre-cancerous lesion to a malignant tumour. Globally, 5-10% of all cancers are attributed to genetic defects and 90-95% to environmental and lifestyle factors such as cigarette smoking, diet, alcohol and physical inactivity.

Additionally, of all cancer-related deaths, almost 25–30% are due to tobacco, 30–35% are linked to diet, about 15–20% are due to infections, and the remaining percentage are due to other factors like radiation, stress, physical activity, environmental pollutants among other prognostic factors (Anand et al 2008). In low and middle income countries, at least 25% of cancers ((Association, 2017)) are also caused by infectious agents including human papilloma virus (cancer of the cervix), hepatitis B and C (cancer of the liver), and helicobacter pylori (cancer of the stomach).

The burden of cancer at the macro and micro level is huge and this is compounded by a severely limited capacity of most low-income countries to provide the necessary health care. Late-stage presentation and inaccessible diagnosis and treatment are common (WHO, 2017). In 2015, only 35% of low-income countries reported having pathology services generally available in the public sector. More than 90% of high-income countries reported treatment services are available compared to less than 30% of low-income countries.

Censoring is a part of observation data's method in the survival analysis. There are consist of right censored, left censored and interval censored. In our study, we will focus more on the interval censoring. According to (Xu, 2012), partly interval censored data consist of exact data and interval censored data. This means that some of the subject event of interest is exactly observed while for other it lays within an interval. There are researchers who still continue using interval censored in their studies as cited by (Xu, 2012). These studies include Zhang (2016) used a class of generalized log-rank test for interval censored failure data and as discussed the work by Razali, (2009) in interval censored data where treating an exact observation as an interval censored observation with very short interval. Elfaki, (2012) also maximized the used of Proportional Hazard Weibull Model in the study of parametric Cox's model for interval censored data of the application for the AIDS study.

In this study, interval censored will be mainly used in order to estimate the survivability of failure rate using Weibull distribution model. The model will be adapted from the model that was developed by Guure, (2013) & Alharphy, (2013) with some modification and different estimator methods. The estimation methods that will be used will consist of analytical methods which will include maximum likelihood estimation.

Some of the available models for survival analysis include non-parametric methods (Kaplan-Meier), semi parametric analysis (Cox's proportional hazard model), and parametric models (Exponential, Weibull, Gamma and lognormal). Considering their flexibility and variety in function and performance, parametric models are of particular interest to many researchers. Exponential Weibull distribution model is an extended version of the Weibull model, which is a flexible, robust model in fitting survival data and can be effectively used based on the data structure (Wahed AS, 2009). To estimate the survival of patients with chronic disease and the prognostic factors in this regard, widely used survival analysis models are often reviewed.

1.1 Review of Cancer and its Causes

Cancer is a general term used for a group of diseases that cause abnormal cells to divide without control and overpass other tissues. In addition, if they expand out of control, cancer can result in death (American Cancer Society, 2014). Based on GLOBOCAN 2012, an estimated 14.1 million new cases of cancer and 8.2 million deaths from cancer occurred in 2012 in both sexes. Estimation of 5-year prevalent cases in 2012 showed that there were 32.5 million people (adult population) alive from both sexes who had a cancer diagnosed during the previous five years (Ferlay et al., 2014).

In Kenya, cancer is estimated to be the third leading cause of death after infectious and cardiovascular diseases. Among the NCD related deaths, cancer is the second leading cause of death accounting for 7% of overall national mortality after cardiovascular diseases (WHO 2014 NCD country pro-le 2014). The annual incidence of cancer is close to 37,000 new cases with an annual mortality of over 28,000. The leading cancers in women are cervix uteri (40.1/100,000), breast (38.3/100,000) and oesophagus (15.1/100,000). In men, prostate (31.6/100,000), Kaposi sarcoma (16/100,000) and oesophagus (20.5/100,000) are the most common cancers (Ferlay et al 2013).

1.2 Review Survival Analysis

Survival analysis is a branch of statistics which analysis of time to events, such as death in biological organisms and failure in mechanical systems. This topic called reliability theory or reliability analysis in engineering, and duration analysis or duration modelling in event history analysis in sociology. Survival analysis attempts to analysis the proportion of a population which will survive past a certain time. The Cox regression model (Cox, 1972) is the most popular method in regression analysis for censored survival data. However, due to the very high dimensional space of the predictors, the standard maximum Cox partial likelihood method cannot be applied directly to obtain the parameter estimates. To deal with the problem of co linearity, the most popular approach is to use the penalized partial likelihood which was proposed by Tibshirani (1995) and is called the least absolute shrinkage and selection operator (Lasso) estimation. In the case of biological survival, unambiguous, but for mechanical reliability, failure well-defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure).

1.3 Review of Possible Prognostic Factors for Cancer Patients Survival

In the reviewed studies by Lee et al.,(2007) in which they analysed retrospectively prognostic factors that affect clinical outcome of breast cancer patients with brain metastases. The overall median survival time was 5.6 months. In their study, 23.1% of patients survive more than 1 year. However, they did not analyse prognostic factors of these long-term survivors, separately. Recent retrospective study evaluated clinical data from 420 patients who had been diagnosed with breast cancer and brain metastasis between 1994 and 2004 at M. D. Anderson Cancer Centre. In this study median follow-up after brain metastasis was 6 months (range 0.7–95.9 months) and the overall median survival was 6.8 months [Altundag K.,2007]. In conclusion, although the survival of patients with breast cancer metastatic to brain is generally poor, there are some long-term survivors such as younger patients with hormone receptor positive histology, good performance status, and single metastatic lesion. More aggressive treatment approach may be considered for these patients. Detailed molecular characterization of brain metastases from breast cancer might lead to more in-depth understanding of those biological abnormalities.

1.4 Weibull Parametric Models

The Weibull Distributions has been widely studied since its introduction in 1951 by Professor Wallodi Weibull (Weibull, 1951). These studies range from parameter estimation; AlFawzan (2000) to diverse applications in reliability engineering especially in Tang (2004) and lifetime analysis in Lawless (2003). The popularity of the distribution is attributable to the fact that it provides a useful description for many different kinds of data, especially in emerging areas such as wind speed and finance (stock prices and actuarial data) in addition to its traditional engineering and medical applications.

In this study, the exponential Weibull model will be used to evaluate the survival of the patients with cancer. This model is in fact a generalized format of the Weibull model, which has two shape parameters (α and β) and a scale parameter (λ). The exponential Weibull function is as follow:

$F(t)=[1-\exp(-\lambda t)^{\alpha}]^{\beta-1}$

$f(t) = \alpha\beta\lambda[\ 1 - (\exp(-\lambda t)^{\alpha}\ ]^{\beta-1} \exp(-\lambda t)^{\alpha}) (-\lambda t)^{\alpha-1}$

For the specified amount of β = 1, the Weibull model will be converted into the Exponential Weibull model; it is anticipated that the flexibility of the model would increase by adding it to a shape parameter.

1.4.1 Gompertz distribution

The non-negative random variable t is said to have a generalized Gompertz distribution (GGD) with three parameters H = (t, α, β) if its cumulative distribution function is given by the following form

$F(t; \alpha, \beta,\lambda) = [1 - e^{\{(\beta/\alpha)(e(\alpha t)-1)\}}]^{\lambda}, \quad t \geq 0, \alpha, \beta > 0.$

The parameter β is a shape parameter. The generalized Gompertz distribution with parameters t, α and β will be denoted by GGD (t; α, β,). The first advantage of GGD is that it has the closed form of its the cumulative distribution function as given in (1).

1.4.2 Exponential Distribution

A new family of distributions, namely the exponential distribution was introduced by Gupta et al. (1998). The family has two parameters (scale and shape) similar to the Weibull or gamma family. Properties of the distribution were studied by Gupta and Kundu (2001). They observed that many properties of the new family are similar to those of the Weibull or gamma family. Hence the distribution can

be used as an alternative to a Weibull or gamma distribution. The two-parameter Weibull and Gamma distributions are the most popular distributions used for analysing lifetime data. The gamma distribution has wide application other than that in survival analysis. However, its major setback is that its survival function cannot be obtained in a closed form unless the shape parameter is an integer. This makes the Gamma distribution less popular than the Weibull distribution, whose survival function and failure rate have very simple and easy-to-study forms.

$G(t) = [F(t)]\nu = \{1-\exp[-(t/\alpha)\beta]\}\nu$, $t \geq 0$.

The density function is

$g(t) = \frac{\beta\nu}{\alpha\beta}\ t^{\beta-1}\, e^{-(t/\alpha)^{\beta}}\{1- e^{-(t/\alpha)\beta^{\nu-1}}$

1.4.3Akaike Information Criterion (AIC)

Akaike Information Criterion (AIC) is a measure of selecting a model from a set of models. AIC estimates the quality of each model, relative to each of the other models. The AIC is given by:

AIC = −2 * log (likelihood) + 2(k) (10)

where k is the number of parameters in model. Thus, k = 1 for the exponential model and k = 2 for the Weibull and gompertz models. AIC can also be calculated using residual sums of squares from regression: AIC = n * log (RSS n) + 2(k) (11)

where n is the number of data points (observations) and RSS is the residual sums of squares. Smaller AIC will indicate a better model fit.

1.4.4. Log-likelihood value

The estimated parameters of three parametric models will be estimated by using maximum likelihood functions as discussed before. The selection of the best fit will depends on the likelihood values of the observe data under three parametric model. The log-likelihood function is given by:

$\log L(\theta; y) = \sum \log f i\ (yi\ ;\ \theta)\ ni=1$ (13)

The model that will gives the highest likelihood value will the best fit model.

1.5 Application to Analysis and Time Default

This study analyses survival of Cancer patients in definite settings that are in real life situations. Weibull distribution Model is chosen. Data collected from MTRH is the used to show how the factors affect the survival of the patient with the model. Weibull model is subjected to the prognostic factors and it showed that the factors were significant with a p value less than 0.05.

Results were obtained as follows

Age at diagnosis and Gender of Cancer Patients

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.0	27.0	36.0	35.8	44.0	53.0 88.0
Female	Male				
112	99				

In total, 211 cancer patients who were referred to Moi Teaching and Referral Hospital, Eldoret-Chandaria Cancer and Chronic Diseases Centre, 2014-2017 were evaluated in this study. Of this, 99 (46.9%) were male while 112 (53.1%) were female. This study shows that female are more prone to cancer disease by 7 % as compared to male counterpart. The mean age at diagnosis of cancer patients was 35.8 where the maximum age 88 and minimum age was 19.

Stage of Infections of Cancer Patients

May not survive beyond six months 79
May survive only for one year 10
Possible to recover 58
Possible to recover fully 64

It was found out that 79 (37.4%) of cancer patients may not survive beyond six months, 10 (4.7%) may survive only for one year, 58 (27.5%) of cancer patients were possible to recover after year and 64 (30.3%) of cancer patients were possibly recover fully after one year.

Histologic Grade of Cancer Patients

Moderately differentiated 78
Poorly differentiated 74
Well differentiated 59

According to the histologic grade, majority of the patients were detected with moderately differentiated tumour 78 (37%), with poorly differentiated tumour 74 (35.1%) and well differentiated tumour 59 (27.9%). This indicates that histologic grade that is well differentiated has 27.9% more likelihood to survive cancer treatment as compared to poorly differentiated and moderately differentiated tumour.

Cancer Type of Cancer Patients

Breast	22
Cervical	14
Colorectal	32
Gastric	8
Hodgkin	21
Kaposi	21
Leukemia	27
Lungs	45
Prostrate	21

According to this study results, the cancer type in these patients was the significant in majority cancer types. For instance Breast 22 (10.4%) ,Cervical 14 (6.6%), Colorectal 32 (15.17%), Gastric 8 (3.8%), Hodgkin Lymphoma 21 (10%), Leukemia 27 (12.8%), Lungs 45 (21.3%), Oesophagus (58.9%), Prostrate 21 (10%) and Kaposi Sarcoma 21 (10%) had significantly affect the survival of a cancer patients.

Comparing the Goodness of fit of Weibull model and other Weibull Related Models.

flexsurvreg (formula = Death ~ Gender + as.numeric(time) + Treatment +
Histologic. Grade + Family. History + Cancer. Type + Bmi + Stages.of.Infections,
data = Cancer, dist. = "gompertz")
N = 211, Events: 191, Censored: 20
Total time at risk: 7554
Log-likelihood = -759.049, df = 23
AIC = 1564.098
flexsurvreg (formula = Death ~ Gender + as.numeric(time) + Treatment +

Histologic. Grade + Family. History + Cancer. Type + Bmi + Stages.of.Infections,
data = Cancer, dist. = "Weibull")
N = 211, Events: 191, Censored: 20
Total time at risk: 7554
Log-likelihood = -759.4498, df = 22
AIC = 1562.9
flexsurvreg (formula = Death ~ Gender + as.numeric(time) + Treatment +
Histologic. Grade + Family. History + Cancer. Type + Bmi + Stages.of.Infections,
data = Cancer, dist. = "exponential")
N = 211, Events: 191, Censored: 20
Total time at risk: 7554
Log-likelihood = -879.9439, df = 21
AIC = 1801.888

From the study results, it indicate that AIC for Gompertz were 1564.098, Weibull were 1562.9while Exponential were 1801.888. Thus, to adequately compare the Weibull and other Weibull related models such exponential and Gompertz the study use Akaike Information Criterion (AIC), as a measure of testing the goodness of fit of an estimated statistical model, to choose the best model knowing that a lower AIC indicates the better fit. The study results show that Weibull model gives the best fit as compared to the other mention parametric models since it had minimal AIC value.

This study results also shows that the likelihood values were as follows; Gompertz were -759.049, Weibull were -759.4498and Exponential were -879.19. The estimated parameters of three parametric models will be estimated by using maximum likelihood functions as discussed before. Thus from the results obtain from the study, it clear that the Gompertz model best fit the cancer data since it has the highest likelihood values.

2. Conclusions and Recommendations

In summary, the results shows that the combine influence of the prognostic factors identified are statistically significant since the overall p value is less than 0.05 ($p<0.05$). This implies that survival of cancer patients will depend on significance effect of prognostic factors such as age at diagnosis, treatment administered, histologic grade, family history, BMI, stage of infection, gender of patient and type of cancer.

Finally, it is recommended that the effects of other prognostic factors such as genetics, lifestyle, and dietary habits on survival time be investigated in further research. The Weibull model could be used to assess the prognostic factors in the healed patients formerly afflicted by cancer.

References

[1] Association, A. M. (2017). Global, Regional, and National Cancer Incidence, Mortality,Years of Life Lost, Years Lived With Disability, and Disability Adjusted Life-years for 32 Cancer Groups, 1990 to 2015. *Clinical Review & Education*, 524-548.

[2] Ferlay J, S. I., Ferlay , J., Soerjomataram , I., Ervik , M., Dikshit, R., Eser , S., . . . GLOBOCAN. (2012). Cancer Incidence and Mortality Worldwide: . *GLOBOCAN*, 11.

[3] Weibull, W. (1951). A Statistical Distribution of wide Applicability. *Journalof Applied Mechanics*, 239-296.

[4] WHO. (2014). World cancer factsheet. *International Agency for Research on Cancer*.

[5] Xu, Z. (2012). Comparing Graphical Method and AModified Method to Fit Weibull Distribution. . *Theses andDissertations Leigh Preserve*, 1149.

[6] Guure, C. B., Ibrahim, N. A., & Adam, M. B. (2012). On Parlty Censored Data with the Weibull Distribution. *ARPN Journal of Engineering and Applied Science*, 1329-1334.