Modeling the Survival of Retrospective Clinical Data from Prostate Cancer Patients in Komfo Anokye Teaching Hospital, Ghana

Asiedu-Addo, Samuel KKwesi
Nwi-Mozu, Isaac
Ali, Mohammed
University of Education, Winneba
Ghana

Abstract

In this study, we model and estimate how long a patient at Komfo Anokye Hospital can live with prostate cancer diseases knowing the status of the cancer. A total of 199 patients (2005-2013) were retrieved. Variables considered were age, treatment and stage of the cancer. AIC was used to select the best parametric model. By comparing different parametric models using AIC, the result indicated that Weibull model fit best. A straight line graph of K-M survival against time {log (-log S (t)) vs. log t)} produced confirmed Weibull assumption. A plot of Kaplan Meier hazard with the Weibull hazard further confirmed the Weibull assumption.

Keywords: Model, Survival, Prostate Cancer, Kaplan Meier

1. Introduction

As science and mathematics progress into technology, abstract thinking becomes real to the physical world. The application of mathematics and statistics in the area of biological and medical fields has gained root in recent studies. See for example Van Erruggen et. al. (2005) and Holbrook & Longest (2013). One area in statistics that makes use of statistical language to explain the characteristics behavior of a system is statistical modeling. As biological and medical problems are viewed by scientists as quantitative, statistical models are formulated to solve those problems. One of the major challenges in medicine is the ability to determine the best treatment for diagnosis. Thus, the best treatment is one that has passed through a statistical test to prove significant effect on a patient. Another challenge is the ability to estimate how long a patient can recover from a treatment or can live with a particular disease. One strategy that has contributed to resolve such problems is survival modeling. Cox (1972) formulated a model that could help medical statisticians to estimate probabilistic future behavior of a patient based on either retrospective cohort or prospective cohort studies. According to Cleves et al. (2010), formulating a model helps to obtain some estimates for use in designing a future survival study. In a study to determine the predictor variables contributing to the survival of prostate cancer in a hospital in Ghana, Asiedu-Addo et. al. (2014) identified the stage of cancer at which a patient is diagnosed as the main contributor to the survival of a patient and recommended that future work must be done to model the survival of a prostate cancer patient. Thus, the aim of this paper was to formulate a model that would help to estimate how long a prostate cancer patient can live with the disease. The model was guided by Murray (1993) recommendation that the art of good modeling must consist of the following: (i) a good understanding and appreciation of the biological problem, (ii) the mathematical model representing idealization of the physical laws taking into account assumption in order to make the model tractable, (iii) appropriate solution must rely on using numerical techniques and (iv) the solution obtained should be consistent with physical intuition and evidence.

2. Data Description

Data for this study consist of all prostate cancer patients diagnosed at Oncology Department at Komfo Anokye Teaching Hospital (KATH) in Kumasi, Ghana within a 9 year period (2005 to 2013). Patient's records were evaluated for the endpoints of either survival or death. In 11 cases out of 210, no information of patient's records was retrievable and hence excluded from the analysis. This approach is similar to the study conducted by Siddiqu et al (2001) in Pakistani breast cancer patients.

Thus, in all 199 prostate cancer patients were used in this study. Generally, in research, data are taken over a finite period of time. Sometimes, time to event may not be observed for most variables in the study population. This result is called censored data. Such observation is predominant in the area of medicine. In medicine, during diagnosis of a particular disease, all patients in the study do not have the same entry time.

Thus, the amount of follow up varies from patient to patient. Also the time to follow up until patients die with the diseases under treatment may not be observed due to a patient still alive before the study ends , a patient is lost to follow-up during the study period or patients withdraws from treatment. Hence, censoring and differential follow-up cause difficulties in the analysis of such data that cannot be handled by the standard statistical methods such as logistic regression or ordinary regression model. Survival analysis is different from the other statistical methods. In survival analysis, dependent variable is always time until occurrence of an event of interest and therefore survival data are generally skewed and thus, the assumption of normality does not hold. In this study, the following category of patients were censored: patients who were alive and disease-free; patients who died from other causes other than prostate cancer; patients who were still alive with the disease as at their last follow up date; and, patients who were lost to follow-up.

3. *Mathematical Formulation*

The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The survival function gives, for every time, the probability of surviving (or not experiencing the event) up to that time. The survival function denoted by S is defined by

$$S(t) = \Pr(T > t)$$

Where, t is some time, T is a random variable denoting the time of death, and "Pr" stands for probability. That is, the survival function is the probability that the time of death is later than some specified time t. The hazard function $\lambda(t)$ gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. For continuous random variables, the hazard function is given by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\Pr(t \le T < t + \Delta t)}{\Pr(T \ge t)} = \frac{S'(t)}{S(t)}.$$

A number of models are available to analyze the relationship of a set of predictor variables with the survival time. Methods include parametric, nonparametric and semi parametric approaches. In this work parametric method was employed.

3.1. Parametric Survival Analysis

Parametric analysis is a branch of statistics that assumes that data has come from a type of probability distribution. In order to perform a parametric survival analysis, we need to identify an appropriate distribution function that characterizes the behavior of the survival times. An accelerated failure time model is a parametric model that considers the linear relationship between the logarithm of survival time and covariates of interests. Furthermore, we assume that the survival time follows a given theoretical probability distribution. Let T_i denote a continuous non-negative random variable representing the survival time and X_i' be the independent variables (covariates) in the model. The analytical structure of the failure model is given by

$$\log(T_i) = \beta X_i' + \varepsilon_i \qquad (1)$$

Where ε_i is an error term that has a suitable probability distribution such as normal, Weibull or gamma and β is a constant to be determined. By exponentiation of (1), we obtain a model for survival time

$$T_i = T_{0i} \exp(\beta X_i') \qquad (2)$$

Where T_{0i} is the exponentiated error term?

The three most widely used parametric models; exponential, Weibull and Log-logistic were considered in this study. The definition of the distributions considered in this study was taken from Table man and Kim (2003) as follows:

3.2 The Exponential Distribution

Exponential distribution has one parameter with hazard function assumed to be constant over time. Consider the density function $f(t)$ given by:

$$f(t) = \lambda \exp(-\lambda t) \qquad (3)$$

The survival function $S(t)$ is given by

$$S(t) = \int_0^\infty f(t)dt \qquad (4)$$

The corresponding hazard function is

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$= \lambda \qquad (5)$$

3.3 The Weibull Distribution

The Weibull model has two parameters and it is obtained if the error term ϵ has an extreme value distribution with the following density:

$$f_\varepsilon(x) = \exp(x - \exp^x), \ -\infty < x < \infty$$

Equivalently, T has the Weibull distribution with the following density:

$$f(t) = \lambda k (\lambda t)^{k-1} \exp(-(\lambda t)^k), \ t \geq 0 \qquad (6)$$

The survival function is:

$$S(t) = \int_0^\infty f(t)dt$$

$$= \exp(-(\lambda t)^k) \qquad (7)$$

The hazard function is given as:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$= \lambda k (\lambda t)^{k-1} \qquad (8)$$

Where λ is the scale parameter and k is the shape parameter. The Weibull distribution is convenient because of its simple form. It includes several hazard shapes and reduces to exponential distribution when $k = 1$

3.4 Logs-Logistic Model

The lifetime T is log-logistically distributed if

$$Y = \log(T)$$

is logistically distributed with location parameter μ and scale parameter γ . Hence, Y is also of the form

$$Y = \mu + \gamma Z$$

Where Z is a standard logistic random variable with density

$$\frac{\exp(z)}{(1+\exp(z))^2}, \ -\infty < z < \infty$$

Equivalently, T has the log-logistic distribution with the following density:

$$f(t) = \lambda k (\lambda t)^{k-1}, \ t \geq 0, \ k > 0, \ \lambda > 0 \qquad (9)$$

We have the survival function as

$$S(t) = \int_0^\infty f(t)dt$$

$$= \frac{1}{1+(\lambda t)^k} \qquad (10)$$

The hazard function is given as:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$\lambda(t) = \frac{\lambda k(\lambda t)^{k-1}}{1+(\lambda t)^k} \qquad (11)$$

4. *Model Selection*

In this study, the Akaike Information Criterion (AIC), which is one of the most widely used in selecting the best competing fitting parametric models, was employed. The AIC is defined in Diez (2012) as:

$$\text{AIC} = -2Log(\text{maximum likelihood}) + k \times p$$

Where p is number of parameters in each model under consideration and k a predetermined constant. The lower the AIC, the better the model.

4.1 Distribution Assumptions

If the Kaplan-Meier survival function can be estimated, then a straight line graph of survival against time indicate a particular parametric distribution. A straight line graphical assumption for each distribution is summarized in Table 1.

Tale 1: Parametric Survival Models Assumption

	Exponential	Weiull	Log-Logistics
Survival S(t)	$S(t) = \exp(-\lambda t)$	$S(t) = \exp(-(\lambda t)^k)$	$S(t) = \frac{1}{1+(\lambda t)^k}$
Assumption	$\log S(t)$ vs. time t	$\log(-\log S(t))$ vs. $\log t$	$-\log\left(\frac{S(t)}{1-S(t)}\right)$ vs. $\log t$

both Korosteleva (2008) and Machin et al. (2006) indicated that Kaplan Meier survival graph can be used to check model appropriateness. Thus, if either exponential or Weibull or log-logistic distribution survival graph behaves similarly to Kaplan Meier survival curve, then that particular model fit the data best. Thus, from this suggestion, the reverse is also true since the compliment of survival is hazard. That is, if any of the distribution behaves similarly to Kaplan Meier hazard graph then that particular distribution fit the data best.

5 *Results and Analysis*

Table 2 provides the AICs for each parametric model under study. The AIC result gave approximate equal values for both the Weibull and Log-logistic models for patient survival which does not depend on the predictor under consideration.

Table 2: Model selection using AIC for no predictor

	Exponential	Weiull	Log-Logistics
AIC: no predictor	561.2192	541.4593	541.4968
AIC: full predictor	561.7859	542.4550	542.4789

The AIC for full predictors in Table 2 indicates that the Weibull distribution is reasonable to model the survival and hazard function.

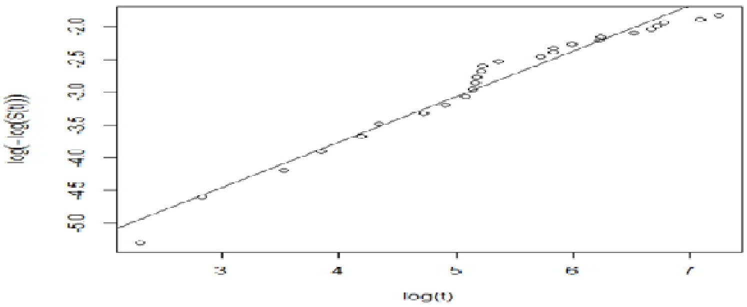

Figure 1: Weibull Assumption graph for no predictor

A graph of $\log t$ against $\log(-\log(S(t)))$ should produce an approximate straight line if the Weibull assumption is true where t is the survival time and $S(t)$ is the Kaplan Meier survival of time. Figure 1 produces an approximate straight line suggesting that the Weibull distribution is
appropriate.

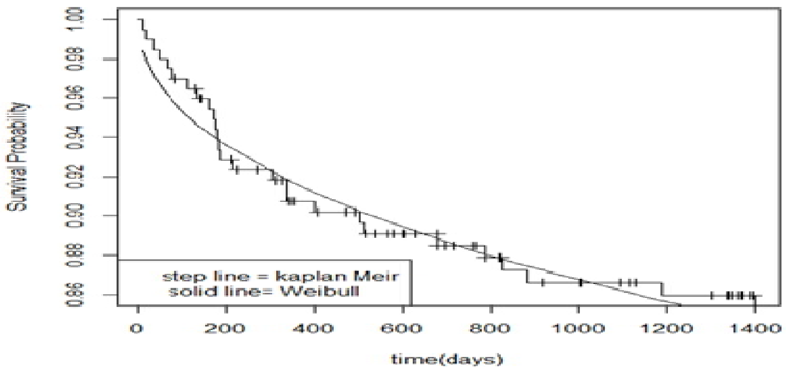

Figure 2: Weibull and K-M survival curve for no predictor

Further assumption is checked if the Weibull model agrees with the Kaplan Meirer survival function. From Figure 2, the parametric curve goes nicely through the Kaplan-Meier estimates. The output in Tale 3 gives the estimation of the Weibull model. Thus, the output of the Weibull distribution in table 3 lead to the survival and the hazard models with no predictor for equation (10 and 11) respectively. The output in Tale 3 gives the estimation of the Weibull model.

Tale 3: Weiull output for no covariate

	Value	Std. Error	Z	P
Intercept	11.330	0.819	13.83	$1.59e^{-43}$
log (scale)	0.721	0.182	3.97	$7.22e^{-05}$

Thus, the output of the Weibull distribution in table 3 lead to the survival and the hazard models with no predictor for equation (10 and 11) respectively.

$$S(t) = \exp(-731\times10^{-11}t^{2.06}) \qquad (10)$$

$$h(t) = 15.06\times10^{-11}t^{1.06} \qquad (11)$$

It can be seen that the hazard rate of a prostate cancer patient increases slowly with time. This means that the longer a patient stay in the study without any influence by covariate, the higher his chance of experiencing the event. The Weibull model gives a close approximation to the Kaplan Meier estimator by comparison. For instance, the probability that a patient will survive in the 250 days is estimated by the Weibull as 0.999956 (approximately 1, higher survival) and that of the Kaplan Meier estimate is about 0.94.

The result also shows that a prostate cancer patient will eventually experience death at the long run. The parametric model become most powerful by most researchers than the non parametric technique in the sense that, the parametric models give an exact estimate than the non-parametric estimator which can only give up to two decimal places. Also, the parametric technique can give a future estimate which the Kaplan Meier non parametric cannot do. For instance, we can estimate the survival and hazard time for a patient in the next 5000 days for parametric model but cannot be estimated by the Kaplan Meier estimator. The model only explain the survival and hazard rate of a patient but does not account for the factors that yield this results. It is therefore, important to investigate if covariate (age, treatment, cancer stage) have significant influence on a prostate cancer survival or death. The investigation can be done in two ways. One is to add possible covariates one at a time and select those which are significant. The second is to start with the full model which includes all possible covariates, and then knock those which show no significant effect on the survival function. For the purpose of this study, the second strategy is adopted. Earlier study by Asiedu-Addo et. al. (2014) gave the output of significant predictor as shown in Tale 4.

Table 4: P-Values for full predictors

Predictor	P-Value
Intercept	1.500655e−28
Age	6.717104e−03
Treatment	2.647685e−01
Stage 2	2.235348e−29
Stage 3	6.523545e−198
Stage 4	$0.000000e^{+00}$

It can be seen from Table 4 that the stage 4 of the prostate cancer appear to be significant but age at which patients were detected prostate cancer was not significant. Patients treatments shows no significant as reported by the Wald test. If the stage of the cancer of a patient is significant it must satisfy the Weibull assumption. The log-log for both stage one and stage two cannot be plotted and so have been omitted. This is because no patient experiences the event for stage one cancer and so its log-log graph cannot be estimated. Only one patient experiences the event and so its log-log graph is also omitted.

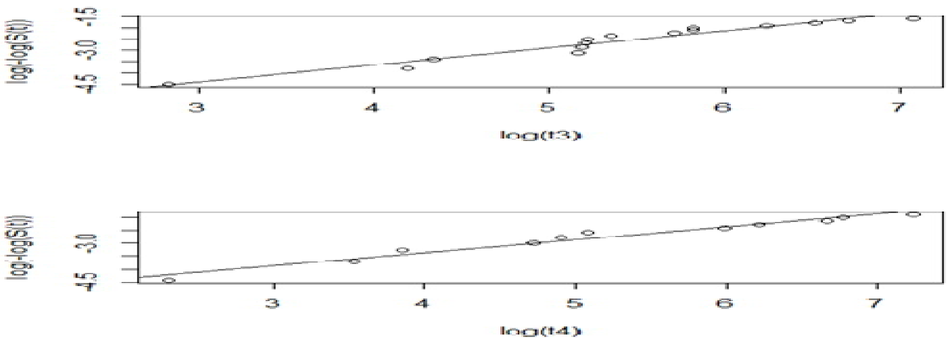

Figure 3: Weibull assumption for prostate cancer stages (stages 3 and 4)

Again a graph of $\log(t)$ against $\log(-\log(S(t)))$ should produce an approximate straight line if the Weibull assumption is true. From figure 3 it can be observed that both graphs give an approximate straight line, suggesting that Weibull assumption is met. The validation of the Weibull assumption is further cheeked with the K-M. graph as show in Figure 4. We observe that both graphs give an approximate straight line, suggesting that Weibull assumption is met. The validation of the Weibull assumption is further checked with the K-M

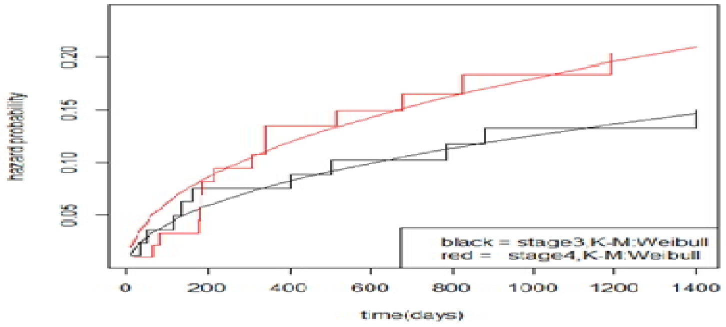

Figure 4: Weibull and K-M hazard graph for prostate cancer stages

It can be observed that the Weibull distribution goes fairly smoothly with the Kaplan-Meier hazard function giving indication to use the Weibull model. According to Korosteleva (2008), it is wise to rerun the model without the insignificant predictors. Based on this suggestion, only the predictor (stage of the cancer) is included in the model and the result of the Weibull output is as shown in Table 5.

Table 5: Weibull estimation

	Value	Std. Error	Z	P
Intercept	47.062	0.823	57.20	0.00e+00
Stage 2	-34.424	2.126	-16.19	6.13e−59
Stage 3	-35.614	0.815	-43.72	0.00e+00
Stage 4	-36.305	0.000	-Inf	0.00e+00
Log (scale)	0.713	0.181	3.93	8.56e−05

The survival and hazard functions for this model are summarized in the Table 6:

Table 6: Weibull survival and hazard model

Survival Stage model	Hazard stage model
$S(t, x_{\text{stage 1}}) = \exp(-20.18\times10^{-43}t^{2.04})$	$h(t, x_{\text{stage 1}}) = 47.17\times10^{-43}t^{1.04}$
$S(t, x_{\text{stage 2}}) = \exp(-63.57\times10^{-13}t^{2.04})$	$h(t, x_{\text{stage 2}}) = 12.97\times10^{-12}t^{1.04}$
$S(t, x_{\text{stage 3}}) = \exp(-7.20\times10^{-11}t^{2.04})$	$h(t, x_{\text{stage 3}}) = 14.69\times10^{-11}t^{1.04}$
$S(t, x_{\text{stage 4}}) = \exp(-29.49\times10^{-11}t^{2.04})$	$h(t, x_{\text{stage 4}}) = 60.16\times10^{-11}t^{1.04}$

The estimated acceleration factor γ comparing the cancer stage 2, stage 3 and stage 4 to stage 1: [(stage 1 vs. stage 2),(stage 1 vs. stage 3),(stage 1, stage 4)] is given by: $\gamma = \exp(\alpha_i)$, where α_i is the coefficient of each cancer stage. The estimated acceleration factor is given by Table 7.

Table 7: Estimated acceleration factor for patients cancer stages

Cancer stage	$\gamma = \exp(\alpha_i)$
Stage 2	$1.122013e^{-15}$
Stage 3	$3.413288e^{-16}$
Stage 4	$1.710208e^{-16}$

From Table 7, it can e seen that the estimated coefficient for stage 2, stage 3 and stage 4 are negatives implying that the survival for patient with prostate cancer stage 2, stage 3 and stage 4 are decreased by a factor of $1.122013e^{-15}$, $3.413288e^{-16}$ and $1.710208e^{-16}$ respectively.

In all, the survival rate decreases with increase in prostate cancer stages. Alternatively, the corresponding hazard can have its interpretation as having a prostate cancer stage 2 accelerated death by a factor of $e^{34.42} = 88.80 \times 10^{15}$ relative to a stage 1 prostate cancer while having a prostate cancer stage 3 accelerated death by a factor of $e^{35.61} = 29.19 \times 10^{16}$ relative to a stage 1 prostate cancer and a prostate cancer stage 4 accelerated death by a factor of $e^{36.31} = 58.78 \times 10^{16}$ relative to a stage 1 prostate cancer. Overall, the hazard of the stages of the cancer is seen to increase in cancer stages relative to stage 1. For instance, the hazard or risk of a patient dying with a prostate cancer in the next 15 years (5475 days) with stage 1, stage 2, stage 3 and stage 4 are 3.18 × 10−38, 10 × 10−8, 1.13 × 10−6 and 4.65 × 10−6 respectively.

6 Discussion and Conclusion

Using the AIC to compare the three most common used distributions, it is found that the Weibull distribution is considered the best fit for the prostate cancer data since the AIC value for the Weibull model is smaller than the Exponential distribution. Again by the law of parsimony, the Weibull distribution was considered over the Log-logistic since is simply used. The Weibull distribution is seen to be very close to the non-parametric Kaplan-Meier distribution, suggesting model accuracy. Table 4 shows patients survival does not depend on age or treatment given. This means a patient of age 89 years with stage 3 prostate cancer will likely have the same survival with patient with age 40 years. The interpretation of the exact value of the co-efficient cancer stage is intuitive. However, the negative sign indicate that, the survivor rate of a patient decreases with increase in prostate cancer stage. The value of stage 1 cancer is given by the Weibull intercept which gave a positive co-efficient sign indicating a perfect survival or low risk.

References

Asiedu-Addo, S. K., Nwi-Mozu, I., Assuah, C. K. & Asumadu, O. R. (2014). Determination of predictor variables contributing to the survival of prostate cancer patients in Komfo Anokye teaching Hospital, Ghana. *International Journal of Applied Science and Technology*. Vol 4 No. 6

Cleves, M., Gould, W., Gutierrez, R. G. & Marchenko, Y. V. (2010). An Introduction to Survival Analysis Using Stata. Stata Press, Third Edition

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, 34(2):187-220. Series B (Methodological).

Diez, D. (2012). Survival Analysis in R. www.openintro.org.

Holbrook, L.T. & Longest, P.W. (2013). Validating CFD predictions of highly localized aerosol deposition in airway models: *in vitro* data and effects of surface properties. *Journal of Aerosol Science*, Vol. 59 pp 6–21.

Korosteleva, O. (2008). Clinical Statistics: Introducing Clinical Trials, Survival Analysis, and Longitudinal Data Analysis. Second edition.

Machin, D., Cheung, Y. B. & Mahesh K. B. (2006). Survival Analysis: A Practical Approach.

Murray, J. D. (1993). Mathematical Biology: Berlin: Springer-Verlag

Siddique, T., Sabih, M., Khan, S. & Salam, A. (2001). A survival analysis of metastatic breast cancer in Pakistani patients. *Journal of Pakistan Medical Association*, 54(67).

Tableman, M. & Kim, J. S. (2004). Survival Analysis Using S: Analysis of Time-to-Event Data. Chapman and Hall/CRC

Van Ertbruggen, C., Hirsch, C. & Paiva, M. (2005). Anatomically based three-dimensional model of airways to simulate flow and particle transport using computational fluid dynamics. *Journal of Applied Physiology* 98(3), 970-980.