

Distribution of Computerized Datasets To Fit Weibull, Log-Logistics and Gompertz Survival Models

Bilkisu Muhammad Bello¹, Farouq Ndamadu Musa²,
and Ibrahim Aliyu Hassan^{3*}

¹Department of Computer Science, Kaduna Polytechnic, Kaduna

²Department of Maths and Statistics, Kaduna Polytechnic, Kaduna

³Department of Environmental Science, Kaduna Polytechnic, Kaduna

*Corresponding author details: Ibrahim Aliyu Hassan;
aliyuibrahim@kadunapolytechnic.edu.ng

ABSTRACT

The lack of detailed and analytic study on the risk factors of Breast cancer in Nigerian presents us with the challenge of not knowing the risk factors distinct to the Nigerian settings. Effectively handling the breast cancer treatment in co-infected patients is delicate. Some challenges to the treatment of breast cancer infection include drug-to-drug interactions between procedures required to treat the cancer cells, thus the hormonal drugs burdens patient's inability to comply with dosage intake thereby making women with breast cancer prone to developing other forms of cancer. The aim of the research was to analyzed distributed dataset to fit Weibull, Log-logistics and Gompertz survival models. The research methodology involved the collection of data cases on oncological study analyzed using descriptive statistics and parametric survival models. The result shows that the Cramer-von Misses (W), the Anderson Darling (A) and the Kolmogorov Smirnov (D) statistics, Log-logistic distribution has greater p-value than other distributions, hence it is the best fit for survival times of a group of patients given chemotherapy treatment. The research concluded that Gompertz model was the best fitted model to survival times of a group of patients given chemotherapy treatment data, while log-logistic distribution was the best fit distribution for the data. It also concluded that Weibull and log-logistic models perform better than Gompertz model in fatigue fracture data, while log-logistic distribution was the best fit for the data and Gompertz model and the distribution was the best fit to nicotine measurements.

Keywords: Gompertz; Weibull; log-logistic; distribution

INTRODUCTION

According to World Health Organization (WHO, 2002), cancers will be ranked among the top 20 causes of death by 2030. Thus, breast cancer will be the most dominant opportunistic disease worldwide. Cancer cells around the thoracic cavity accelerates the development of breast cancer, by increasing spread of the infection. Though some risk factors are known, these identified risk factors are not entirely exhaustive since other risk factors are being discovered continuously. The lack of detailed and analytic study on the risk factors of Breast cancer in Nigerian presents us with the challenge of not knowing the risk factors distinct to the Nigerian settings. Effectively handling the breast cancer treatment in co-infected patients is delicate. Some challenges to the treatment of breast cancer infection include drug-to-drug interactions between procedures required to treat the cancer cells, thus the hormonal drugs burdens patient's inability to comply with dosage intake thereby making women with breast cancer prone to developing other forms of cancer. Other variables include Remission Times of Bladder Cancer Patients, Myelogenous leukemia data, Survival times of a group of patients given chemotherapy, Fatigue Fracture data, and Nicotine measurements respectively. Due to the complex nature of breast cancer and other forms of cancers, it becomes difficult for medical practitioners to prioritize treatment of either over the other.

The key to mitigating and addressing this menace is by making the risk factors familiar through research. This will make medical practitioners and care givers prioritize the kind of treatment needed for that same reason. There is a knowledge gap in the breast cancer control and preventive measures associated with its infection.

LITERATURE REVIEW

The review provides conceptual sources and empirical studies on Comparative Study on The Performance of Weibull, Log-Logistic and Gompertz Survival Models on Oncological Data.

Conceptual Framework

The occurrence of survival (or time-to-event) data is commonplace in medical research, where interest lies in the time it takes from a given baseline, for an event of interest to occur, and the factors that are associated with it. For example, this could be the effect of a treatment on the time to death since diagnosis of cardiovascular disease.

The two main approaches to survival analysis, are the semi-parametric approach of Cox (1972), and fully parametric approaches, assuming such distributions as the exponential or Weibull, for example (Collett, 2003).

The Cox model does not assume any functional form for the baseline hazard function, whereas a parametric approach assumes a specific shape, estimated as part of the model. Both allow us to investigate the influence that risk factors have on the rate of disease or mortality, for example. In this research we would want to concentrate on the parametric approach to survival analysis, in particular, deriving a general algorithm to simulate survival data under more biologically plausible scenarios to better assess both methods used in practice, and novel models.

Empirical review

Cancer is the name given to a collection of related diseases. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. It is one of the leading causes of death in the world and represents a tremendous burden on patients, families and societies. There were 12.7 million new cancer cases in worldwide, of which 5.6 million occurred in developed countries and 7.1 million in developing countries. The corresponding estimates for total cancer deaths were 7.6 million 2.8 million in developed countries and 4.8 million in developing countries. There were an estimated 4.9 million new cases and 0.266 million global deaths from cervical cancer accounting for 7.5% of all female cancer deaths. Cervical cancer is one of the leading causes of death in the world and represents a tremendous burden on patients, families and societies. It is estimated that over one million women worldwide currently have cervical cancer; most of them have not been diagnosed or have no access to treatment that could cure them or prolong their lives Survival data is a term used for describing data that measure a time to the occurrence of a given event of interest. In this study the event of interest is survival time of cervical cancer patients from the day of diagnosis. One of the major aims of this analysis was to assess the survival of women with cervical cancer using various parametric frailty models. Kaplan and Meier obtained one important development in non-parametric methods. The non-parametric methods work well for homogeneous samples; they do not determine whether certain variables are related to the survival times. The Cox PH model has the restriction that proportional hazards assumption holds with time-fixed covariates; and it may not be appropriate in many situations and other modifications such as stratified Cox model or Cox model with time-dependent variables are required times.

DATA ANALYSIS AND PRESENTATION

(1) Descriptive Statistics

TABLE 1: One Sample test about the distribution of dataset for Myelogenous.

Model	1	O W	A	Kolmogorov-Smirnov test	
				D	p-value
Weibull		0.05147	0.40237	0.66589	< 2.2e-16
Log-logistics		0.3082024	1.957631	0.46635	< 2.2e-16
Gompertz		0.1523498	0.9057371	0.10096	0.2547

Source: Field Survey, 2022

Table 1 presents the Cramer-von Misses (W), the Anderson Darling (A) and the Kolmogorov Smirnov (D) statistics, it is observed Gompertz distribution has greater p-value than other distributions, implying that the Gompertz distribution is the best fit for Myelogenous leukemia data.

Although the Cox regression model is the most favorable employed technique in survival analysis, parametric models do have a number of benefits.

Theoretical Review

Survival analysis is one of the primary statistical methods for analyzing data on time to an event which is a data that has an end point. That is the time when an event occurs such as death but may include other kinds of events which can either be positive or negative such as time to discharge or time to recovering, heart attack, device failure, etc. These data analysis is important for different legal proceedings which include estimating cost of future medical care, years of life lost, evaluating product reliability, and assessing drug safety and so on. This branch of empirical science entails gathering and analyzing data on time until failure or death. Survival analysis includes different types of data analysis including life table analysis, time to failure methods, and time to death analysis.

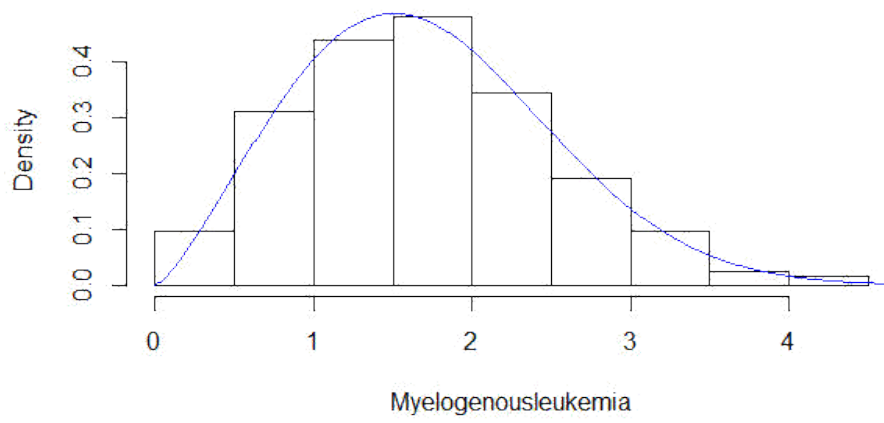
RESEARCH METHODOLOGY

Research Design: Data cases on oncological study is used in this study, obtained from internet sources and publications. Descriptive Statistics of dataset is performed using mean, median, mode, variance, skewness, and kurtosis. Parametric survival models are used in the analysis. The models are Weibull, Log-logistics and Gompertz models, the models are chosen because of their similarities in order to have better basis for comparison and also have differences that will cater to the situation where the other one fails. The models are fitted to the data with the view to find the best fit. R statistical package was used for analyzing the data.

Survival analysis

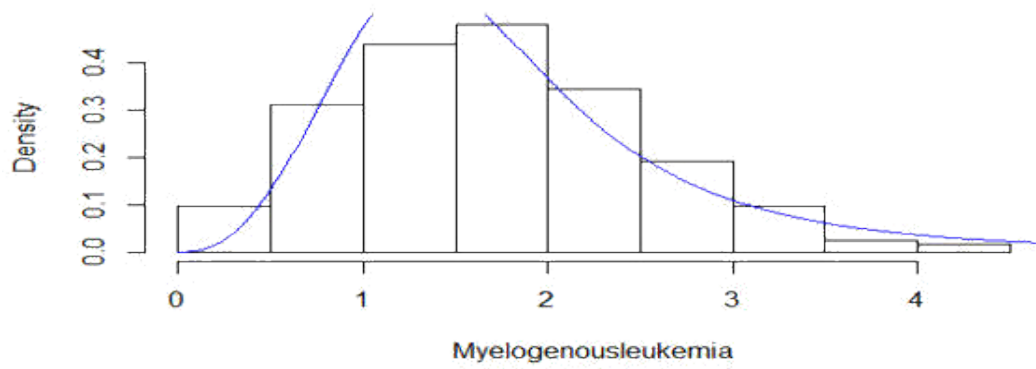
A survival model is used to analyze time-to-event data and to generate estimates, referred to as survival curves, that show how the probability of the event occurring changes over time. In many life situations, as time progresses, certain events are more likely to occur. The survival models help decision-makers to form better estimates than guessing about the expected timing of certain events. There are three types of survival model: parametric model, semi-parametric model, and non-parametric model, parametric model is the model selected for this design.

Histogram of Myelogenousleukemia



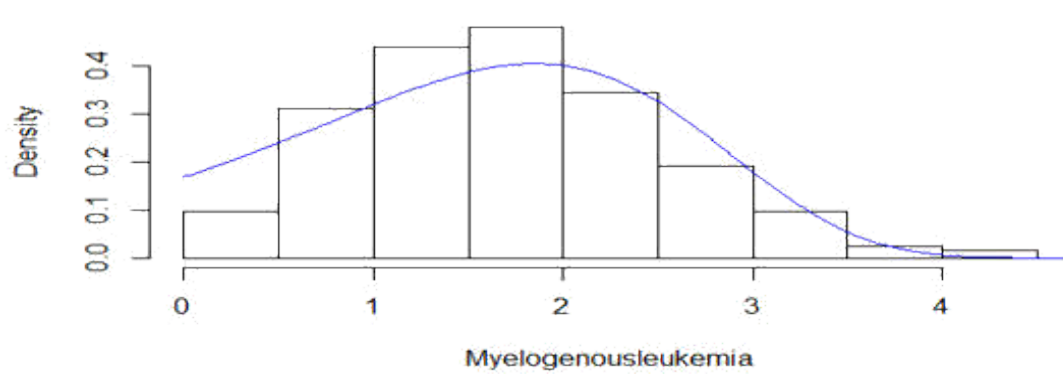
(A) Weibull

Histogram of Myelogenousleukemia



(B) Log-logistics

Histogram of Myelogenousleukemia



(C) Gompertz

FIGURE 1: Fitted curve of the three distributions for myelogenous leukemia data.

(2) Survival Analysis on Survival Times of a Group of Patients Given Chemotherapy Treatment

TABLE 2: MLE's and Information Criteria of models for survival times of a group of patients given chemotherapy treatment.

Model	$\hat{\alpha}$	$\hat{\beta}$	AIC	CAIC	BIC	HQIC	LL
Weibull	0.833042	1.487036	124.384	124.6697	127.9973	125.731	60.19198
Log-logistic	0.833042	1.487036	124.384	124.6697	127.9973	125.731	60.19198
Gompertz	0.258314	1.974204	121.377	121.6630	124.9906	122.7243	58.68863

Source: Field Survey, 2022

Table 2 shows the maximum likelihood estimates of dataset 3 with 3 with different model fit, all the information criteria and log-likelihood of the models, Gompertz model has smallest value in all the information criteria, indicating that Gompertz model is the best fitted model to survival times of a group of patients given chemotherapy treatment data.

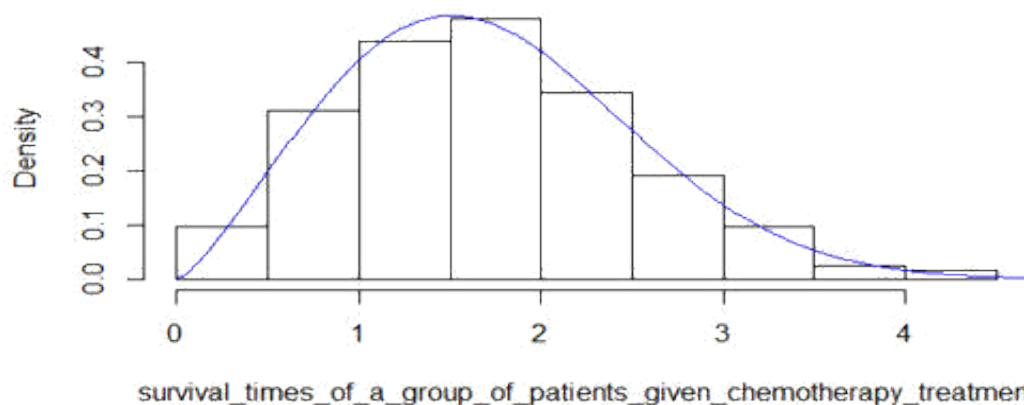
TABLE 3: One Sample test about the distribution of dataset for survival times of a group of patients given chemotherapy treatment (Dataset3).

Model	W	A	Kolmogorov-Smirnov test	
			D	p-value
Weibull	0.13391	0.87063	0.27477	0.00171
Log-logistics	0.08041826	0.5569135	0.087438	0.8519
Gompertz	0.1132596	0.742632	0.16673	0.146

Source: Field Survey, 2022

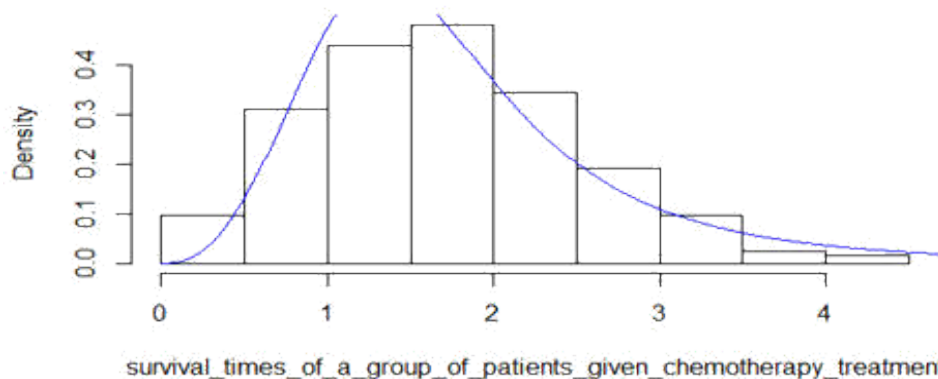
Table 3 presents the Cramer-von Misses (W), the Anderson Darling (A) and the Kolmogorov Smirnov (D) statistics, Log-logistic distribution has greater p-value than other distributions, hence it is the best fit for survival times of a group of patients given chemotherapy treatment.

im of survival_times_of_a_group_of_patients_given_chemotherap



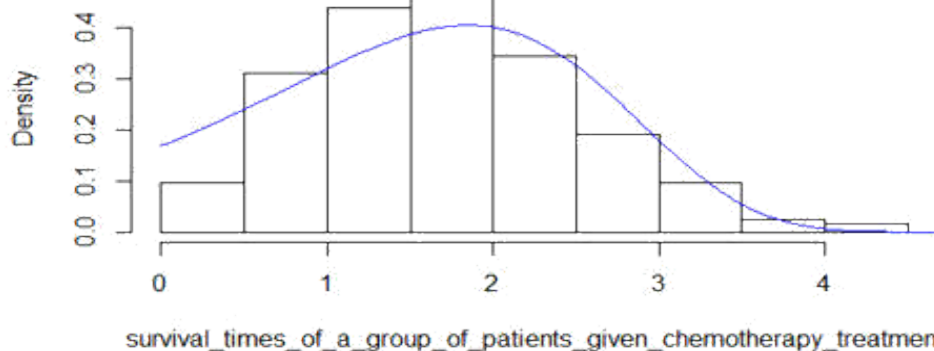
(A) Weibull

im of survival_times_of_a_group_of_patients_given_chemotherap



(B) Log-logistic

im of survival_times_of_a_group_of_patients_given_chemotherap



(C) Gompertz

FIGURE 2: Fitted curve of the three distributions on survival times of a group of patients given chemotherapy treatment.

(3) Survival Analysis on Fatigue Fracture Data

TABLE 4: MLE's and Information Criteria of models for Fatigue Fracture data (Dataset4).

Model	$\hat{\alpha}$	$\hat{\beta}$	AIC	CAIC	BIC	HQIC	LL
Weibull	1.565866	1.963127	252.9542	253.1185	257.6156	254.8171	124.4771
Log-logistic	1.565866	1.963127	252.9542	253.1185	257.6156	254.8171	124.4771
Gompertz	0.197519	1.880304	256.6414	256.8058	261.3029	258.5044	126.3207

Source: Field Survey, 2022

Table 4 shows the maximum likelihood estimates of dataset 1 with 3 with different model fit, all the information criteria and log-likelihood of the models. Weibull and Log-logistic models have smallest value in all the information criteria, indicating that Weibull and log-logistic models perform better than Gompertz model in fatigue fracture data.

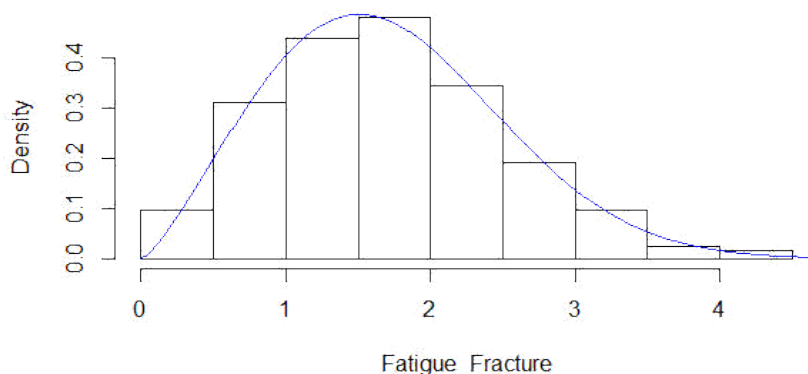
TABLE 5: One Sample test about the distribution of dataset for fatigue fracture (Dataset4).

Model	W	A	Kolmogorov-Smirnov test	
			D	p-value
Weibull	0.2666667	1.529223	0.22099	0.0009683
Log-logistics	0.1526425	0.9209887	0.099912	0.4073
Gompertz	0.2328952	1.331934	0.11856	0.2178

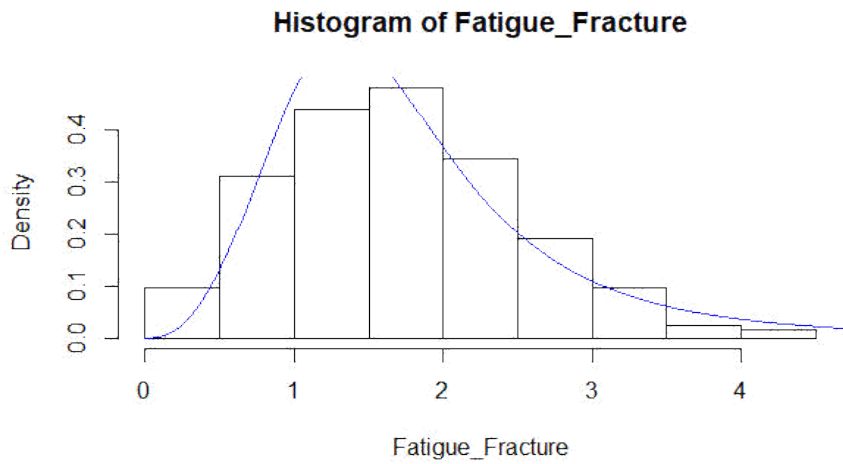
Source: Field Survey, 2022

Table 5 presents the Cramer-von Misses (W), the Anderson Darling (A) and the Kolmogorov Smirnov (D) statistics. Log-logistic distribution has greater p-value than other distributions indicating that the log-logistic distribution is the best fit for fatigue fracture data.

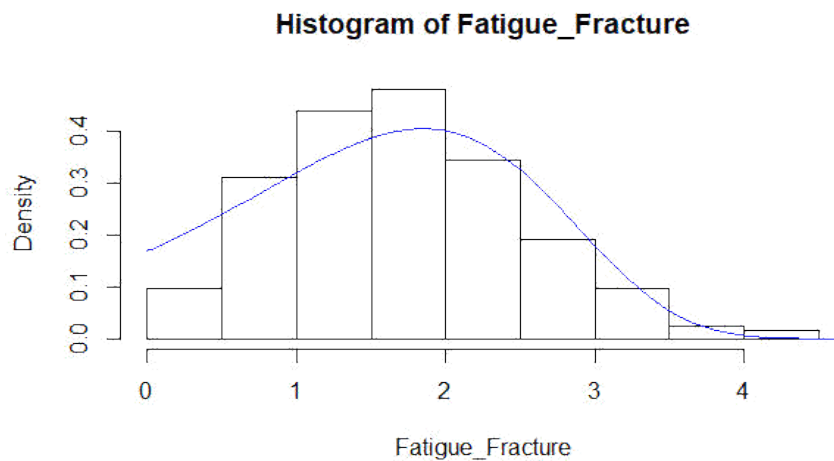
Histogram of Fatigue_Fracture



(A) Weibull



(B) Log-logistic



(C) Gompertz

FIGURE 3: Fitted curve of the three distributions on Fatigue Fracture Data.

(6) Survival Model for Nicotine Measurements

TABLE 6: MLE's and Information Criteria of models for nicotine measurements (Dataset5).

Model	$\hat{\alpha}$	$\hat{\beta}$	AIC	CAIC	BIC	HQIC	LL
Weibull	0.869319	1.972623	462.84	462.875	470.5329	465.9033	229.42
Log-logistic	0.869319	1.972623	462.84	462.875	470.5329	465.9033	229.42
Gompertz	1.796779	0.217613	284.43	284.469	292.1264	287.4968	140.22

Source: Field Survey, 2022

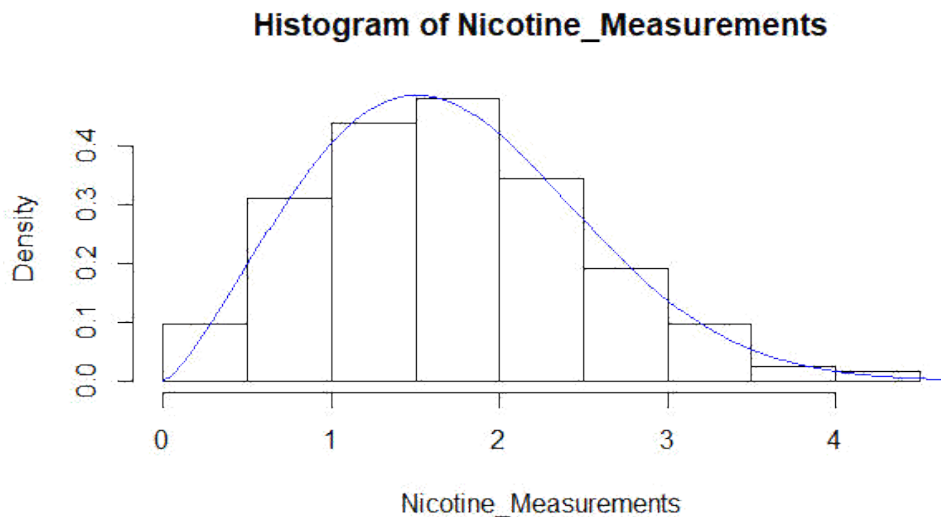
Table 6 shows the maximum likelihood estimates of dataset5 with 3 with different model fit, all the information criteria and log-likelihood of the models. Gompertz model has smallest value in all the information criteria, indicating that Gompertz model is the best fitted model to nicotine measurements data.

TABLE 7: One Sample test about the distribution of dataset for nicotine measurement (Dataset5).

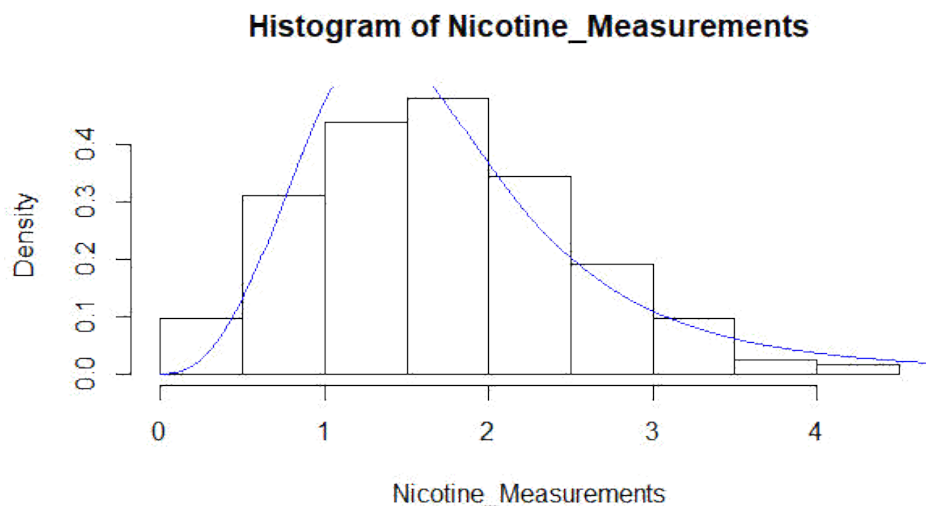
Model	W	A	Kolmogorov-Smirnov test	
			D	p-value
Weibull	0.6492512	3.774388	0.22814	4.441e-16
Log-logistics	1.582193	9.350919	0.25369	< 2.2e-16
Gompertz	0.4903747	2.944429	0.17034	3.807e-09

Source: Field Survey, 2022

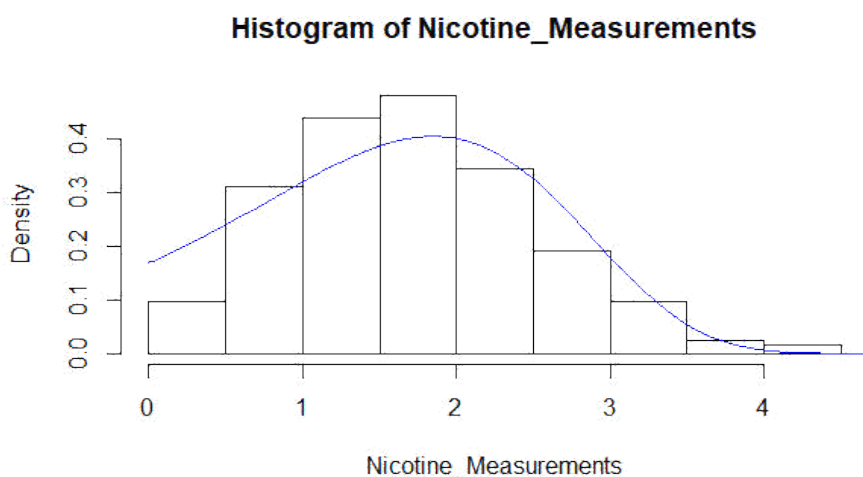
Table 7 presents the Cramer-von Misses (W), the Anderson Darling (A) and the Kolmogorov Smirnov (D) statistics, Gompertz distribution has greater p-value than other distributions, therefore Gompertz distribution is the best fit for nicotine measurements.



(A) Weibull



(B) Log-logistic



(C) Gompertz

FIGURE 4: Fitted curve of the three distributions on nicotine measurements.

CONCLUSION

Based on the analysis carried out, it was concluded that Gompertz model was the best fit in the oncological data, followed by the Log-logistic model, Weibull and Log-logistic model behave similarly on the dataset.

RECOMMENDATION

The research has shown that most of the research conducted on oncological data (cancer related cases) has shown that there is scarcity of research on related cases on survival model that can best be used to model cancer related data.

Therefore, the research was able to identify a parametric that can best be used to model cancer data known as Gompertz model was the best used model on the research. The research will enable other researcher such as medical personnel to model or know the best model on cancer related cases.

REFERENCES

- [1] Abouammoh, A. M., Abdulghani, S. A., and Qamber, I. S. (1994). On partial orderings and testing of new better than renewal used classed. *Reliable eng. Syst. Saftey*, 43, 37-41.
- [2] Andrews, D. F., Herzberg, A. M. (2012). *Data: a collection of problems from many fields for the student and research worker*. Springer Science Business Media.
- [3] Feigl and Zelen (1965): Survival times (in months) of sample of 101 patients with Advanced Acute Myelogenous leukemia CBNJournal of Applied Statistics Vol. 8 No. 2 (December, 2017)
- [4] Gieser P.W., Chang M.N., Rao P.V., Shuster J.J. and Pullen J. (1998). Modelling cure rates using the Gompertz model with covariate information. *Stat Med* 17(8):831-839.<http://www.ftc.gov/reports/tobacco> or <http://pw1.netcom.com/rdavis2/smoke.html>
- [5] P. E. Oguntunde, A. O. Adejumo & K. A. Adepoju (2016). Assessing the Flexibility of the Exponentiated Generalized Exponential Distribution, *Pacific Journal of Science and Technology*, 17(1), 49-57
- [6] ferlay et. al. (2014). *Cancer Epidemiol Biomarkers prev*2014; 23:963-966 published online.
- [7] Gould, C., Froese, T., Barrett, A. B., Ward, J., and Seth, A. K. (2014). An extended case study on the phenomenology of spatial form synaesthesia. *Front. Hum. Neurosci.* 8:433. doi:10.3389/fnhum.2014.00433

APPENDICES**APPENDIX I: DATASETS****Dataset 1**

Remission_Times_of_Bladder_Cancer_Patients (source)
(0.08,2.09,3.48,4.87,6.94,8.66,13.11,23.63,0.20,2.23,3.52, 4.98,6.97,9.02,13.29,0.40,2.26,3.57,5.06,7.09,9.22,13.80,2 5.74,0.50,2.46,3.64,5.09,7.26,9.47,14.24,25.82,0.51,2.54,3. 70,5.17,7.28,9.74,14.76,26.31,0.81,2.62,3.82,5.32,7.32,10. 06,14.77,32.15,2.64,3.88,5.32,7.39,10.34,14.83,34.26,0.90 ,2.69,4.18,5.34,7.59,10.66,15.96,36.66,1.05,2.69,4.23,5.41, 7.62,10.75,16.62,43.01,1.19,2.75,4.26,5.41,7.63,17.12,46. 12,1.26,2.83,4.33,5.49,7.66,11.25,17.14,79.05,1.35,2.87,5. 62,7.87,11.64,17.36,1.40,3.02,4.34,5.71,7.93,11.79,18.10, 1.46,4.40,5.85,8.26,11.98,19.13,1.76,3.25,4.50,6.25,8.37,1 2.02,2.02,3.31,4.51,6.54,8.53,12.03,20.28,2.02,3.36,6.76,1 2.07,21.73,2.07,3.36,6.93,8.65,12.63,22.69)

Dataset2**Myelogenous leukemia**

(0.03, 8.882, 41.118, 6.151, 17.303, 0.493, 9.145, 45.033, 6.217, 17.664, 0.855, 11.48, 46.053, 6.447, 18.092, 1.184, 11.513, 46.941, 8.651, 18.092, 1.283, 12.105, 48.289, 8.717, 18.750, 1.48, 12.796, 57.401, 9.441, 20.625, 1.776, 12.993, 58.322, 10.329, 23.158, 2.138, 13.849, 60.625, 11.48, 27.73, 2.5, 16.612, 0.658, 12.007, 31.184, 2.763, 17.138, 0.822, 12.007, 32.434, 2.993, 20.066, 1.414, 12.237, 35.921, 3.224, 20.329, 2.5, 12.401, 42.237, 3.421, 22.368, 3.322, 13.059, 44.638, 4.178, 26.776, 3.816, 14.474, 46.48, 4.441, 28.717, 4.737, 15, 47.467, 5.691, 28.717, 4.836, 15.461, 48.322, 5.855, 32.928, 4.934, 15.757, 56.086, 6.941, 33.783, 5.033, 16.48, 6.941, 34.211, 5.757, 16.711, 7.993, 34.77, 5.855, 17.204, 8.882, 39.539, 5.987, 17.237)

Dataset3**survival_times_of_a_group_of_patients_given_chemoth erapy_treatment**

(0.047, 0.115, 0.121, 0.132, 0.164, 0.197, 0.203, 0.260, 0.282, 0.296, 0.334, 0.395, 0.458, 0.466, 0.501, 0.507, 0.529, 0.534, 0.540, 0.641, 0.644, 0.696, 0.841, 0.863, 1.099, 1.219, 1.271, 1.326, 1.447, 1.485, 1.553, 1.581, 1.589, 2.178, 2.343, 2.416, 2.444, 2.825, 2.830, 3.578, 3.658, 3.743, 3.978, 4.003, 4.033)

Dataset4**Fatigue Fracture**

(0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.03.903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960)

Dataset5**Nicotine_Measurements**

(1.3, 1.0, 1.2, 0.9, 1.1, 0.8, 0.5, 1.0, 0.7, 0.5, 1.7, 1.1, 0.8, 0.5, 1.2, 0.8, 1.1, 0.9, 1.2, 0.9, 0.8, 0.6, 0.3, 0.8, 0.6, 0.4, 1.1, 1.1, 0.2, 0.8, 0.5, 1.1, 0.1, 0.8, 1.7, 1.0, 0.8, 1.0, 0.8, 1.0, 0.2, 0.8, 0.4, 1.0, 0.2, 0.8, 1.4, 0.8, 0.5, 1.1, 0.9, 1.3, 0.9, 0.4, 1.4, 0.9, 0.5, 1.7, 0.9, 0.8, 0.8, 1.2, 0.9, 0.8, 0.5, 1.0, 0.6, 0.1, 0.2, 0.5, 0.1, 0.1, 0.9, 0.6, 0.9, 0.6, 1.2, 1.5, 1.1, 1.4, 1.2, 1.7, 1.4, 1.0, 0.7, 0.4, 0.9, 0.7, 0.8, 0.7, 0.4, 0.9, 0.6, 0.4, 1.2, 2.0, 0.7, 0.5, 0.9, 0.5, 0.9, 0.7, 0.9, 0.7, 0.4, 1.0, 0.7, 0.9, 0.7, 0.5, 1.3, 0.9, 0.8, 1.0, 0.7, 0.7, 0.6, 0.8, 1.1, 0.9, 0.9, 0.8, 0.8, 0.7, 0.7, 0.4, 0.5, 0.4, 0.9, 0.9, 0.7, 1.0, 1.0, 0.7, 1.3, 1.0, 1.1, 1.1, 0.9, 1.1, 0.8, 1.0, 0.7, 1.6, 0.8, 0.6, 0.8, 0.6, 1.2, 0.9, 0.6, 0.8, 1.0, 0.5, 0.8, 1.0, 1.1, 0.8, 0.8, 0.5, 1.1, 0.8, 0.9, 1.1, 0.8, 1.2, 1.1, 1.2, 1.1, 1.2, 0.2, 0.5, 0.7, 0.2, 0.5, 0.6, 0.1, 0.4, 0.6, 0.2, 0.5, 1.1, 0.8, 0.6, 1.1, 0.9, 0.6, 0.3, 0.9, 0.8, 0.8, 0.6, 0.4, 1.2, 1.3, 1.0, 0.6, 1.2, 0.9, 1.2, 0.9, 0.5, 0.8, 1.0, 0.7, 0.9, 1.0, 0.1, 0.2, 0.1, 0.1, 1.1, 1.0, 1.1, 0.7, 1.1, 0.7, 1.8, 1.2, 0.9, 1.7, 1.2, 1.3, 1.2, 0.9, 0.7, 0.7, 1.2, 1.0, 0.9, 1.6, 0.8, 0.8, 1.1, 1.1, 0.8, 0.6, 1.0, 0.8, 1.1, 0.8, 0.5, 1.5, 1.1, 0.8, 0.6, 1.1, 0.8, 1.1, 0.8, 1.5, 1.1, 0.8, 0.4, 1.0, 0.8, 1.4, 0.9, 0.9, 1.0, 0.9, 1.3, 0.8, 1.0, 0.5, 1.0, 0.7, 0.5, 1.4, 1.2, 0.9, 1.1, 0.9, 1.1, 1.0, 0.9, 1.2, 0.9, 1.2, 0.9, 0.5, 0.9, 0.7, 0.3, 1.0, 0.6, 1.0, 0.9, 1.0, 1.1, 0.8, 0.5, 1.1, 0.8, 1.2, 0.8, 0.5, 1.5, 1.5, 1.0, 0.8, 1.0, 0.5, 1.7, 0.3, 0.6, 0.6, 0.4, 0.5, 0.5, 0.7, 0.4, 0.5, 0.8, 0.5, 1.3, 0.9, 1.3, 0.9, 0.5, 1.2, 0.9, 1.1, 0.9, 0.5, 0.7, 0.5, 1.1, 1.1, 0.5, 0.8, 0.6, 1.2, 0.8, 0.4, 1.3, 0.8, 0.5, 1.2, 0.7, 0.5, 0.9, 1.3, 0.8, 1.2, 0.9)