
Prediction of Survival of HIV/AIDS Patients from Various Sources of Data Using AFT Models

Markos Abiso Erango^{*}, Ayele Taye Goshu

School of Mathematical and Statistical Sciences, Hawassa University, Awasa, Ethiopia

Email address:

markos.erango73@gmail.com (M. A. Erango)

^{*}Corresponding author

To cite this article:

Markos Abiso Erango, Ayele Taye Goshu. Prediction of Survival of HIV/AIDS Patients from Various Sources of Data Using AFT Models. *Science Journal of Applied Mathematics and Statistics*. Vol. 5, No. 4, 2017, pp. 127-133. doi: 10.11648/j.sjams.20170504.11

Received: April 28, 2017; **Accepted:** May 16, 2017; **Published:** July 7, 2017

Abstract: The aim of this paper is to predict and compare the survival of HIV/AIDS patients under ART follow-up in three different hospitals in Ethiopia. Three data sets with total 1304 patients were considered. Three parametric accelerated failure time distributions: lognormal, loglogistic and Weibull are used to analyze, predict and compare survival probabilities of the patients. The results indicate that the empirical hazard rates of the three data sets reveal maximal peaks. The patients from Arba Minch hospital seems to have highest event intensity. The AFT loglogistic model is selected to best fit to each of the data sets. Different covariates except TB infection status are found to affect patients' survival at each of the hospitals. Patients with TB infection at baseline tend to have shorter survival time as compare to one with no TB infection, with significant differences of survive time between the two groups. Patients under follow-up at Shashemene hospital tend have consistently highest survival probabilities in both TB positive and negative groups. Patients from Bale Robe hospital tend to have longest survival time, while those from Arba Minch hospital have shortest survival time. Patients with bedridden status have the shortest survival time. The AFT-loglogistic is recommended in modelling time-to-event data considered in this study. The results are unique to each hospital implying that patients' care and intervention needs to be specific.

Keywords: Accelerated Failure Time, HIV/AIDS, Prediction, Survival Analysis

1. Introduction

Antiretroviral therapy (ART) has improved the survival of HIV/AIDS patients. Quality of life of the patients has been generally improved worldwide. In resource-poor countries, in particular, the ART has reduced mortality rate among treated HIV/AIDS patients [1], [2]. There are many circumstances in which both a repeatedly-measured biomarker outcome and the elapsed time to an event are collected on each individual in a medical study. These observed biomarker series are frequently important health indicators that represent the progression of a disease. Such data typically have additional features and complications associated with them, including the presence of treatment group indicators and baseline covariates, measurement error in the biomarkers, and right censoring of the event time with the possibility of dependent censoring [3-4]. A study by [4] has demonstrated joint modelling of longitudinal observation of CD4 counts and

time-to-death using AFT models under Bayesian settings. They analyzed two of the data sets with various models and found out interesting results on how covariates and shared frailty affect survival outcome of the patients.

In this study, three data sets are analyzed using AFT models, namely, lognormal, loglogistic and Weibull distributions with classical estimation approach. The purpose is to compare these models and predict survival probabilities in the three different population of patients.

2. Data Description

Three data sets are analyzed in this data sets Data 1, Data 2, Data 3. The three data sets are collected with similar settings. Descriptions of the data given here below.

Data 1: The first data set on CD4 cell counts and survival time are obtained from Shashemene Referral Hospital. A total of 354 random sample of HIV/AIDS patients who had been under ART follow-up from January 2006 to December 2012

in the hospital were considered.

Data 2: The second data set on CD4 cell counts and survival time are obtained from Bale Robe General Hospital. A total of 400 random sample of HIV/AIDS patients who had been under ART follow-up from January 2008 to March 2015 in the hospital were considered.

Data 3: The third data set is obtained from Arba Minch General Hospital for 550 adult HIV/AIDS patients under follow-up from September 2006 to August 2016. Patients who had at least three CD4 measurements after the first report of HIV diagnosis are eligible for the study.

Predictors considered are: sex, age, functional status, TB infection status at baseline, awareness about ART, condom use, OIS - number of opportunistic infections, weight, WHO stage, type of employment, tobacco use, alcohol use, drug use.

3. AFT Models

Survival models are important statistical methods to describe and analyze the time-to-death data of HIV/AIDS patients. An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times in a particular group. In summarizing survival data, the two common functions applied are the survivor function and the hazard rate functions.

The basic quantity employed to describe time-to-event process is the survival function, the probability of an individual surviving beyond time t . Moreover, the distribution of survival time is characterized by three functions: the probability density function, the survivorship function, and the hazard function [5] - [9].

In survival analysis, an accelerated failure time (AFT) model is a parametric model that provides an alternative to the commonly used proportional hazards models for the analysis of survival time data. Under AFT models, we measure the direct effect of the explanatory variables on the survival time instead that of the hazard [3], [4], [6].

Let $T_i = \min(t_i, c_i)$ be the observed time for the i^{th} subject, where t_i is the time-to-event and c_i represents the censoring time which is assumed independent of t_i where $\delta_i = 1$ if the event is observed and $\delta_i = 0$ otherwise. Let $X_{2i} = (X_{21}, X_{22} \dots X_{2p})$ be a vector of p covariates. The corresponding log-linear form of the AFT model with respect to time is given generally as:

$$\log(T_i) = \mu + X_i^T \alpha + \sigma_\epsilon \epsilon_i \quad (1)$$

where μ is intercept, α is the vectors of unknown coefficients, σ_ϵ is the scale parameter and ϵ_i is a random variable assumed to have a particular distribution $f(t)$. Three AFT models used in this study are as defined in [4]:

Lognormal distribution with survival and hazard functions:

$$f(t) = \frac{\sqrt{\tau}}{\sqrt{2\pi}t} \exp\left(-\frac{\tau}{2}\{\ln(t) - \ln(\mu)\}^2\right)$$

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{1/\sqrt{\tau}}\right)$$

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

The parametric link to the covariates and random effects is:

$$T_i \sim \text{lognormal}(\lambda_i(t), \sigma_\epsilon^2), \log(\lambda_i(t)) = X_i^T \alpha + \sigma_\epsilon \epsilon_i \quad (2)$$

Loglogistic distribution with survival and hazard functions:

$$f(t) = \frac{\lambda p t^{p-1}}{(1 + \lambda t^p)^2}$$

$$S(t) = \frac{1}{1 + \lambda t^p}$$

$$h(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p}$$

The parametric link to the covariates and random effects is:

$$T_i \sim \text{Loglogistic}(\rho, \lambda_i(t)), \log(\lambda_i(t)) = -\rho(X_i^T \alpha + \sigma_\epsilon \epsilon_i) \quad (3)$$

Loglogistic and lognormal distributions have hazard rate functions that are non-monotonic that is increasing to reach a peak and then declining over time [3], [4], [6]

Weibull distribution with survival and hazard functions are:

$$f(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$$

$$S(t) = \lambda p t^{p-1}$$

$$h(t) = \exp(-\lambda t^p)$$

The parametric link to the covariates and random effects is:

$$T_i \sim \text{Weibull}(\rho, \lambda_i(t)), \log(\lambda_i(t)) = -\rho\{X_i^T \alpha + \sigma_\epsilon \epsilon_i\} \quad (4)$$

Under the AFT model, we can make prediction of the future survival probabilities given history data of the patients, which is given as:

$$P(s|t, \alpha) = \Pr(T \geq s | T > t, \alpha) \quad (5)$$

Comparison of Models

Model comparison and selection are among the most common tasks of statistical practice. The most commonly used methods of selection include Akaike information and likelihood based criteria. In this study, the AIC criterion used to compare the parametric models, defined as:

$$\text{AIC} = -2LL + 2p \quad (6)$$

where LL is the log-likelihood, p is the number of parameters in the model. Smaller value of AIC suggests a better model in fitting it to the data [10].

4. Results and Discussion

The objective of this study is compare survival probabilities of HIV/AIDS patients under ART follow-up in three different hospitals using three accelerated failure time models. Data 1 are obtained from Shashemene Referral Hospital, Data 2 from Bale Robe General Hospital and Data 3 from Arba Minch General Hospital.

For Data 1, we consider 354 HIV/AIDS patients among which 58.5% are females and 41.5% are males. The average baseline CD4 cell counts is estimated to be 156.58 per mm^3 of blood with standard deviation of 92.54. Among the patients, 94.1% are censored while 5.9% are dead. The average survival time of the patients is estimated to be 48.69 months with standard deviation of 21.27. For Data 2, the average number of baseline CD4 counts is about 177.6 per mm^3 of blood sample with standard deviation of 104.8. Among the 400 sample of patients considered 88.5% of them are censored while 11.5% are dead. For Data 3, the average number of baseline CD4 counts was 162.47 per mm^3 of blood with standard deviation of 102.51. The results of the analysis showed that from the 550 patients included in the study, about 24.4% of them were dead while 75.6% were censored, 65.1% of the patients has less than 200 CD4 count at baseline and 34.9% were greater than or equal to 200 CD4 counts at baseline. About 67.3% are in working, 28.5% in ambulatory and 4.2% in bedridden functional status. Regarding WHO disease stage, 14.9%, 32.7%, 42.5% and 9.8% of them are respectively in stage 1, stage 2, stage 3 and stage 4.

4.1. The Empirical Hazard Rates

The empirical hazard rate estimates of the three data sets are plotted in Figure 1. They show non-monotonic behaviors of the hazard rates, showing suitability of lognormal and loglogistic models instead of Weibull in analyzing these data sets. The maximum hazard rate for Data 1 is estimated to be 0.001508 at time 56.186 months or 4.682 years. The maximum hazard rate for Data 2 is 0.001971 at time 29.678 months or 2.473 years. For data set from Arba Minch General Hospital, the maximum hazard rate is about 0.008438 at time 64.7955 months or 5.399 years. The results indicate that the patients under follow-up at Arba Minch General Hospital

might have higher event intensity as compared to those at both Shashemene and Bale Robe Hospitals.

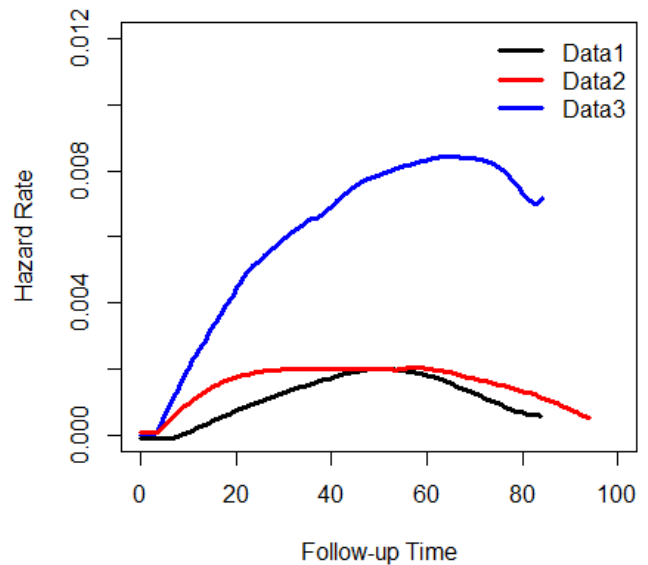


Figure 1. Plots of Empirical Hazard Rates Estimated from Three Data Sets.

4.2. Comparison of Survival Probabilities

To compare the event experiencing time of two or more groups the survival function used of the groups is good indication. To obtain a closer look at estimate of the survival timethe Kaplan-Meier estimation technique was used. The estimated survival functions in Figure 2 show declining of the survival probabilities over time (in each of the data sets). The pattern of survivorship function lying above another means the group defined by the upper curve had a better survival than the group defined by the lower curve.

Comparison by Sex

In case of sex group, there is no significant difference until peak point of hazard rate. However, we can observe a slight difference in survival probability between women and men after peak point of the hazard rates.

The log rank test is a non-parametric test for comparing two or more independent survival curves. To compare the survival probabilities between sex categories, we employ the log-rank statistical test with hypothesis: $H_0: S_F(.) = S_M(.)$ vs $H_1: S_F(.) \neq S_M(.)$.

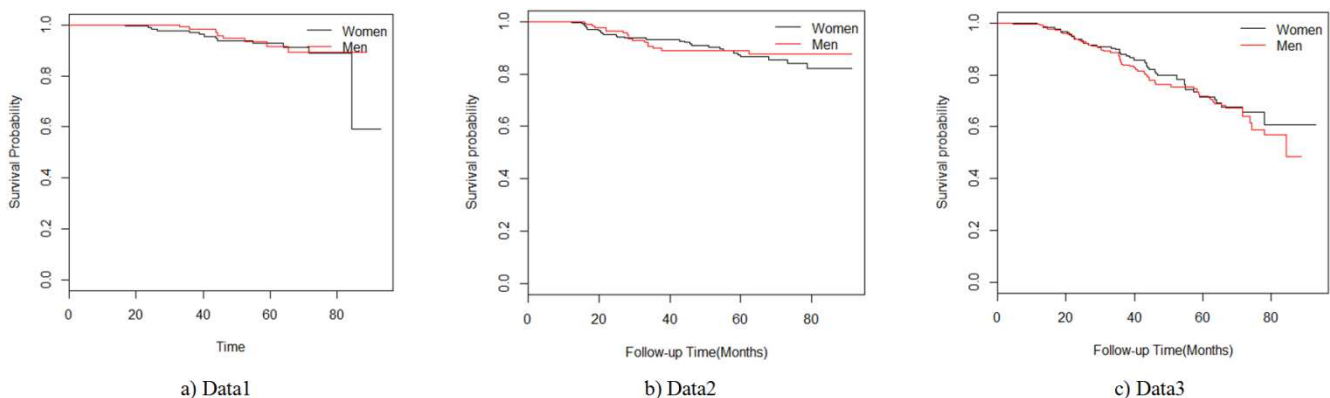


Figure 2. Kaplan-Meier Estimates of the Survival Functions for Women and Men Patients.

The analysis show that the log-ranktest for sex of Data 1 is: ($\chi^2=0.3$ with 1 df, $p=0.607$). The log-ranktest for sex of Data 2: ($\chi^2=0.2$ with 1 df, $p=0.672$), and that for Data 3 is: ($\chi^2=0.5$ with 1 df, $p=0.491$). At 5% level of significance, there is not sufficient evidence to reject the null hypothesis that survival probability of women patients equal to that of

males. Both gender have same survival probabilities.

Comparison TB Status

In all the cases, a patient with TB infection has lower survival probability as compared to that with no TB infection. The difference is significant in all the cases.

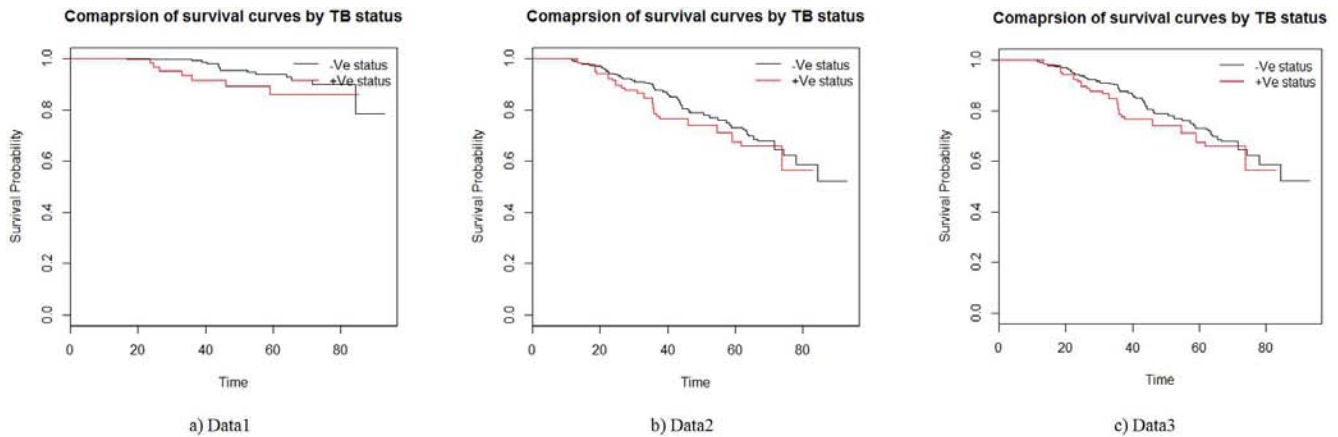


Figure 3. Survival Curves for TB Status at Baseline of the Three Data Sets.

4.3. Analysis for Model Comparison

Analysis of the three data sets for model comparison are given in Table 3. Estimates of total AIC for the models AFT-Lognormal, AFT-Loglogistic, and for AFT-Weibull are displayed. In case of Data 1, the AFT-Weibull best fits to the data set. However, Weibull hazard rate has monotonic behavior that does not match with the empirical hazard rate that is revealing hump-shaped. Thus the Loglogistic model, with the next smallest AIC, is suggested for analyzing Data 1.

In case of Data 2, the AFT-Loglogistic model has the smallest total AIC. Fortunately, the hazard rate of loglogistic distribution behaves like that of the empirical hazard rate. Hence it is considered to be the final model for Data 2.

In the case of Data 3, we consider loglogistic as best fitting model, for same reason given in case of Data 1. Thus the AFT loglogistic model is selected to fit to all the three data sets.

Table 1. Model Comparison Results of Three AFT Models for Three Data Sets.

AIC	Lognormal	Loglogistic	Weibull
Data 1	291.9916	289.4597	286.1394
Data 2	515.4254	515.2777	532.1755
Data 3	1567.883	1561.938	1557.65

4.4. Analysis of the Data Sets Using Loglogistic Model

Case Study of Data 1. See the results in Table 2. Under the AFT loglogistic model, the intercept term is significant at 5% significant level. The covariates having significant effects on survival times of patients are TB status at baseline, awareness about ART, condom use and opportunistic infection. But age, functional status, sex, weight, employment, WHO stages, tobacco, drug and alcohol use have no statistically significant effects on survival of HIV/AIDS patients.

Table 2. Results of Analysis of Data 1 using Loglogistic Model.

Parameter	Estimate	Std. Error	Z	P
Intercept	4.087	0.688	5.939	0.000*
Sex	0.335	0.226	1.485	0.138
Age	-0.010	0.012	-0.890	0.374
Functional status	-0.164	0.144	-1.134	0.257
TB status	-0.505	0.216	-2.336	0.020*
Awareness ART	-0.319	0.135	-2.359	0.018*
Condom use	0.867	0.258	3.354	0.001*
OIS	0.236	0.105	2.248	0.025*
Weight	0.010	0.010	1.000	0.317
WHO	0.005	0.124	0.042	0.966
Employment	0.225	0.225	1.000	0.317
Tobacco use	-0.010	0.242	-0.041	0.967
Alcohol use	-0.201	0.284	-0.707	0.479
Drug use	-0.064	0.308	0.207	0.836
Log(scale)	-1.117	0.177	-6.323	0.000

Case Study of Data 2. See the results in Table 3. Under the AFT loglogistic model, the intercept term is significant at 5% significant level. The covariates having significant effects on survival times of patients are age, TB status at baseline, functional status, condom use, weight, employment and Tobacco use. However, opportunistic infection, sex, awareness about ART, WHO stage, drug and alcohol use are not significant.

Table 3. Results of Analysis of Data 2 using Loglogistic Model.

Parameter	Estimate	Std. Er	Z	p
Intercept	4.635	0.682	6.793	0.000*
Sex	-0.038	0.157	-0.245	0.807
Age	-0.017	0.006	-2.834	0.005*
Functional status	0.177	0.086	2.059	0.040*
TB status	-0.570	0.128	-4.466	0.000*
Awareness ART	-0.328	0.196	-1.672	0.095
Condom use	0.338	0.083	4.093	0.000*

Parameter	Estimate	Std. Er	Z	p
OIS	-0.032	0.044	-0.724	0.469
Weight	0.031	0.010	2.967	0.003*
WHO	0.060	0.100	0.602	0.547
Employment	-0.335	0.105	-3.197	0.001*
Tobacco	-0.338	0.156	-2.171	0.030*
Alcohol use	0.090	0.191	0.474	0.636
Drug use	-0.317	0.197	-1.606	0.108
Log(scale)	-1.089	0.121	-9.024	0.000

Case Study of Data 3. See the results of the analysis in Table 4. Under the loglogistic model, the intercept term is significant at 5% significant level. The covariates having significant effects on survival times of HIV/AIDS are opportunistic infection, TB status at baseline, Functional status and condom use.

Table 4. Results of Analysis of Data 3 using Loglogistic Model.

Parameter	Value	Std. Er	Z	P
Intercept	3.989	0.370	10.791	0.000*
Sex	-0.094	0.098	-0.958	0.338
Age	-0.005	0.005	-1.031	0.303
Functional status	0.339	0.106	3.193	0.001*
TB status	-0.480	0.123	3.913	0.000*
Awareness ART	-0.066	0.071	-0.927	0.354
Condom use	4.068	0.000	0.000	0.000*
OIS	0.381	0.062	6.156	0.000*
Weight	0.009	0.005	1.574	0.115
WHO stage	0.070	0.061	1.148	0.251
Employment	0.088	0.116	-0.758	0.448
Log(scale)	-0.835	0.072	-11.588	0.000

4.5. Prediction of Survival Probabilities Using Covariates

Here we predict survival probabilities and survival times for the HIV/AIDS patient given their TB and functional status. TB status is categorized as with TB infection or no infection. Functional status is categorized in to three levels: working, ambulatory and bedridden.

a) Predictions of Survival Times based on TB Status

TB status is an important factor of survival of the patients. The results of predictions of survival time for HIV/AIDS patient given TB status are displayed in Figures 4 & 5.

In case of Data 1, when the TB status at baseline is positive, the maximum expected survival time of the HIV/AIDS patients is 640 months or 53 years with 0.02 probability. But the TB status at baseline is negative, the maximum expected survival time is 864 months or 72 years with 0.02 probability. This implies that a patient with no TB infection at baseline may survive additional 19 years compared to the HIV patients with TB infection.

In case of Data 2, when a patient gets TB infection at baseline, the maximum expected survival time is 668 months or 56 years with 0.02 probability. For one with no TB infection, the maximum expected survival time is about 933 months or 78 years with 0.02 probability - with 22 additional years.

For Data 3, when a patient is with positive TB status at baseline, the maximum expected survival time is about 552 months or 46 years. For a patient with negative TB status, it is about 636 months or 53 years with survival probability of 0.02 - with 7 additional years.

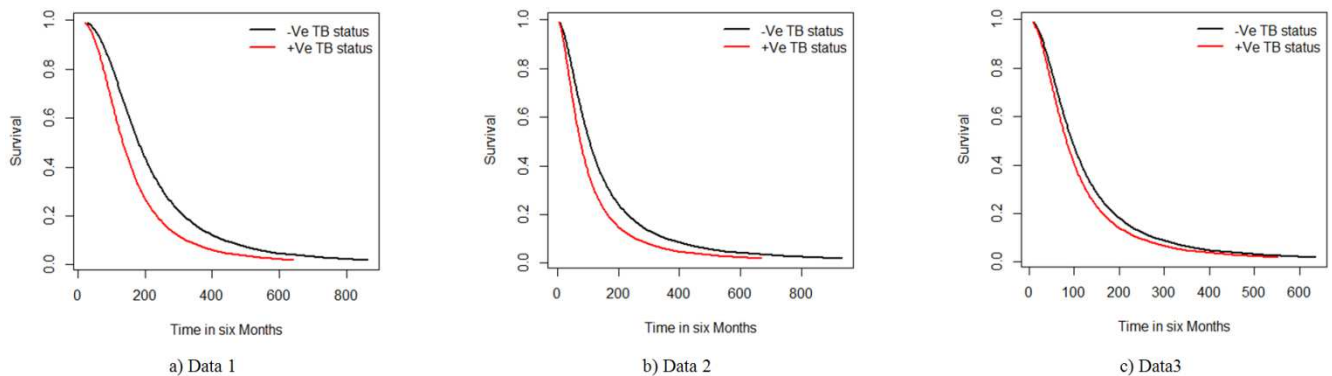


Figure 4. Predictions of Survival Probabilities of Patients based on TB Status - Separately Plotted for each Data Set.

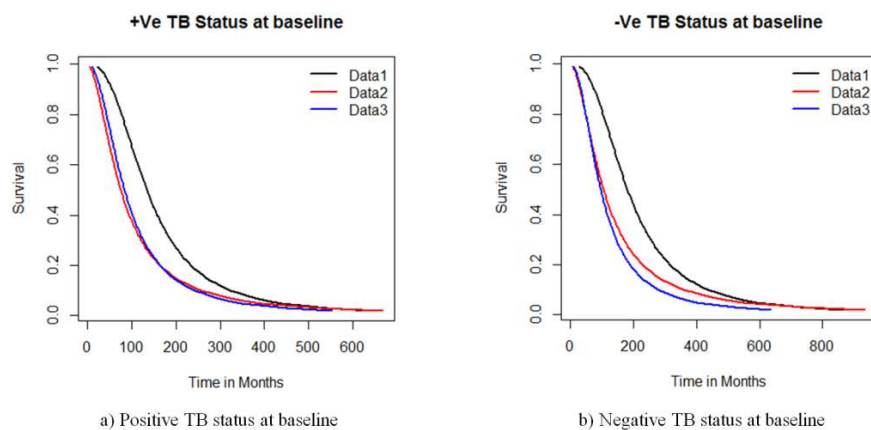


Figure 5. Prediction of Survival Probabilities of Patients based on TB Status - Jointly Plotted for All Data Sets.

Figure 5 displays same results with all data sets together. We can observe from the prediction of future survival that patients with under follow-up at Shashemene hospital will have consistently higher survival as compared to those in other hospitals. Patients with under follow-up at Bale Robe hospital will the longest survival time, while those from Arba Minch hospital will have shortest survival time.

b) Predictions based on Functional Status

Figure 6 displays prediction results with functional status of patients. The results show that the patients in working status will have higher survival probabilities and survival time than those with ambulatory and bedridden status. Patients with bedridden status are those in serious sickness and will have the shortest survival time.

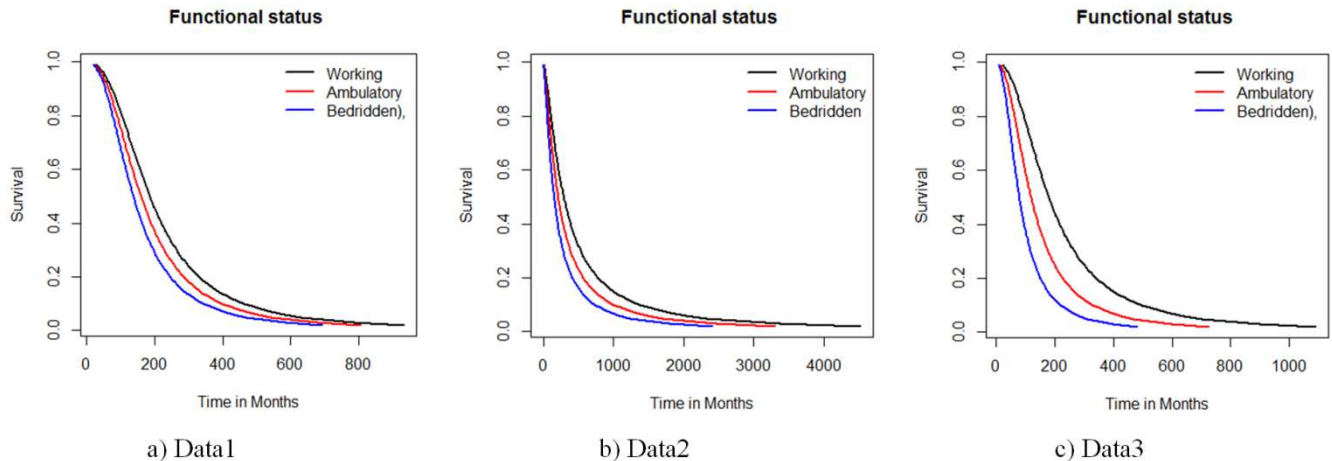


Figure 6. Predictions of Survival Probabilities based on Functional Status.

5. Conclusions

In this study, we aim to predict and compare the survival status of HIV/AIDS patients under ART follow-up at three different hospitals using three AFT models: lognormal, loglogistic and Weibull models. Three data sets are considered.

Based on model comparison and behavior of hazard rate functions, the AFT loglogistic model is selected to best fit to all the data sets. Under this model, the results of analysis show that different covariates affect the survival of the patients from different hospitals. For patients under follow-up at Shashemene General Hospital (Data 1), TB status at baseline, awareness about ART, condom use, and opportunistic infection significantly affect their survival time at 5% significant level. For patients under follow-up at Bale Robe hospital (Data 2), the predictors age, functional status, TB status at baseline, weight, type of employment, and tobacco use are found significant. For the third data set where patients are under follow-up at Arba Minch hospital, the predictors functional status, TB status at baseline, condom use, and opportunistic infection significantly affect survival time of patients.

When predicting future survival status based on TB status at baseline, a patient with TB infection at baseline has short survival time as compared to one with no TB infection. This is consistent in all the data sets. The estimated differences of mean survival time for a patient with no TB infection are 19, 22 and 7 years for the three hospitals, respectively. Patients under follow-up at Shashemene hospital have consistently highest survival probabilities in both TB positive and negative groups. Moreover, patients from Bale Robe hospital

tend to have longest survival time, while those from Arba Minch hospital have shortest survival time. With respect to functional status, patients in working status have highest survival probabilities and longest survival time. Patients with bedridden status have shortest survival time.

AFT-loglogistic is recommended in modelling time-to-event data considered in this study. Patients from different hospitals reveal different results in this study. Thus the results are unique to each hospital implying that patients' care and intervention needs to be specific.

References

- [1] Andinet W., San, M. (2010). Determinants of Survival in Adult HIV Patients on Antiretroviral Therapy in Oromia, Ethiopia. *Global Health Action*, 3:5398.
- [2] Azin, S. A. (2010). An Overview on the 2008 UNAIDS Report on the Global AIDS Epidemic. *Iranian Journal of Epidemiology*, 6(2), 56–58.
- [3] Yu, M., Taylor, J., and Sandler, H. (2008). Individualized Prediction in Prostate Cancer Studies using a Joint Longitudinal-Survival-Cure Model. *Journal of the American Statistical Association*, 103, 178-187.
- [4] Markos A. Erango, Ayele T. Goshu, Gemed B. Buta and Ahmed H. Dessiso (2017). Bayesian Joint Modeling of Survival of HIV/AIDS Patients using Accelerated Failure Time Data and Longitudinal CD4 Cell Counts. *British Journal of Medicine & Medical Research*, 20(6), 1-12. DOI: 10.9734/BJMMR/2017/32123.
- [5] Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis*. John Wiley and Sons, Inc., New York.

- [6] Collett, D. (2003). Modelling Survival Data in Medical Research. Chapman and Hall, London.
- [7] Lee, E. T., and Wang, J. W. (2003). Statistical Methods for Survival Data Analysis. Third Edition. Wiley, New York.
- [8] Hayat, E. A, Suner, A., Uyar, B., Dursun, Ö., Orman, M. N., Kitapçıoğlu, G. (2010). Comparison of Five Survival Models: Breast Cancer Registry Data from Ege University Cancer Research Center. *TurkiyeKlinikleri J Med Sci*, 30(5).
- [9] Klein, J. P. and Moeschberger, M. L. (2003). Survival Analysis Techniques for Censored and Truncated Data. Second Edition. Statistics for Biology and Health, Springer LLC.
- [10] Burnham, Kenneth P. and David R. Anderson. 2002. Model Selection and Multi model Inference: A Practical Information-Theoretical Approach. 2nd ed. New York: Springer.