On the Estimation of Survival of HIV/AIDS Patients on Antiretroviral Therapy Using NPMLE Method: An Application to Interval Censored Data

Gurprit Grover\textsuperscript{1}, Rabindranath Das\textsuperscript{2}, Prafulla Kumar Swain\textsuperscript{1,*}, Barnali Deka\textsuperscript{1}

\textsuperscript{1}Department of Statistics, University of Delhi, Delhi, India
\textsuperscript{2}Department of Statistics, University of Burdwan, West Bengal, India

Abstract The main objective of this paper is to estimate the survival of HIV/AIDS patients who are undergoing Antiretroviral Therapy in an ART centre, Delhi. Non Parametric Maximum Likelihood Estimation NPMLE (E-M) for interval censoring and KM survival plot for left, right and mid-point imputation have been used to estimate the survival of these patients. It has been observed that the mid-point imputed survival plot has a very similar and consistent pattern as obtained by NPMLE (E-M) method. Considering these mid-point imputed value as right censored data, Cox PH model and Accelerated Failure time Model (AFTM) have been applied to identify the effects of prognostic factors like age, sex, mode of transmission, baseline CD4 cell count, hemoglobin, baseline weight and smoking habits on survival of the patients. The Akaike Information Criterion (AIC) has been employed to compare the efficiency of these models and Cox-Snell residual is used to test the proportionality assumption.

Keywords AIDS, ART, CD4 Cell Count, Non Parametric Maximum Likelihood Estimation, Cox-PH Model, Accelerated Failure Time Model

1. Introduction

The widespread use of Antiretroviral Therapy (ART), prognosis of HIV infected patients has significantly improved and mortality due to AIDS related causes has substantially reduced\cite{1, 2, 3}. With the advent of free ART programme in 2004 in India, a dramatic decline of HIV incidence has been observed during the last decade\cite{4}. Thus it is imperative to know the survival of HIV/AIDS patients on ART, so that appropriate strategies can be devised to prevent new infection.

However in HIV/AIDS studies, where time to event data are collected by assessing patients in periodic follow up visits. In such cases, the event cannot be observed exactly (i.e onset of HIV infection, incubation period of AIDS after HIV infection, death time); however it is known to happen within some interval, thus such observed events are interval censored. In HIV dynamics every patient is supposed to visit ART centre after four weeks, but actual visit time vary from patient to patient and also time between visits vary. The patients may visit ART centre at time that is convenient to them rather than scheduled time. Therefore the failure event (i.e death) is considered as interval censored survival data. Survival time $T_i$ (say) of a patient lies in the interval $(L_i, R_i)$ that is the last available visit date and end of the study.

Several approaches have been proposed for the estimation of survival function for the interval censored data; Turnbull\cite{5}, De Gruttola and Lagakos\cite{6} and Sun\cite{7, 8}. More recently Kim\cite{9}, Grover and Shakeri\cite{10}, Grover and Banerjee\cite{11} estimated survival of HIV-1 infected children for doubly and interval censored data.

In this paper we have estimated the survival function of HIV/AIDS patients on ART by using Non Parametric Maximum Likelihood Estimation (NPMLE) method for interval censored data. The NPMLE is computed by using the E-M algorithm of Turnbull (1976) with the polishing algorithm of Gentleman and Geyer\cite{12}. We have also estimated survival function by using imputation technique. Due to lack of appropriate estimation procedure and non availability of statistical software for interval censored data, we have adopted conventional imputation approach to convert interval censored data to right censored data, for which the standard techniques are available. Further, we have compared the estimates obtained by NPMLE and imputation method.

In this paper we have estimated the survival function of HIV/AIDS patients on ART by using Non Parametric Maximum Likelihood Estimation (NPMLE) method for interval censored data. The NPMLE is computed by using the E-M algorithm of Turnbull (1976) with the polishing algorithm of Gentleman and Geyer\cite{12}. We have also estimated survival function by using imputation technique. Due to lack of appropriate estimation procedure and non availability of statistical software for interval censored data, we have adopted conventional imputation approach to convert interval censored data to right censored data, for which the standard techniques are available. Further, we have compared the estimates obtained by NPMLE and imputation method.

Kaplan-Meier plot for left, right and mid-point imputation has been used to estimate the survival of patients. It has been observed that the mid-point imputed survival plot has a very similar and consistent pattern as obtained by NPMLE.
maximizing \( L(P) \) with respect to \( P \), say \( \hat{P} \), by using EM algorithm[14]. Starting with initial estimate \( \bar{P} \),

E-step, for each \( i \)

\[
\hat{P}(t_i) = \frac{P_i(t_j | t_j \epsilon [(L_i,R_i)])}{\sum_{k \epsilon [(L_i,R_i)]} \hat{P}(t_k)}
\]

M-step: update \( P \) for each \( j \)

\[
\hat{P}(t_j) = \frac{\sum_{i=1}^{n} P_i(t_j)}{n}
\]

Iterate the procedure until convergence.

2.2. Imputation Approach and K-M Survival Plot

For simplicity, imputation techniques have been used to handle this interval censored data. After imputation the interval censored data will be treated as complete or right censored data, then standard statistical methods can be performed on the imputed data set (Hsu et al.[13],[16, 17] have used mid-point imputation for interval censored AIDS infection time data. Since the survival time \( T_i \) is known to lie in the interval \( \{(L_i, R_i)\}, \ i = 1, 2, ... n \) for \( i^{th} \) patient. we use the midpoint (m_i) of the observed interval as the hypothetical time \( T \) and the explanatory variables X is given by

\[
\log(T_i) = \beta'X + \epsilon
\]

where the \( \epsilon \)'s denote the distinct imputed exact failure times, d_i's and n_i's are the death and number at risk at each of the \( t_j \)'s respectively.

2.3. Cox-PH Model

The Cox proportional hazard function[18] for the survival time \( T \) and the explanatory variables \( X \) is given by

\[
h(t/X) = h_0(t) \exp(\beta'X)
\]

where \( h_0(t) \) is the baseline hazard. The explanatory variables act multiplicatively on the baseline hazard that is completely unspecified, and the regression parameter \( \beta \) in the model can be estimated by maximizing the partial likelihood function;

\[
L(\beta) = \prod_{j=1}^{m} \frac{\exp(\beta'X_j)}{\sum_{i \epsilon R(t_j)} \exp(\beta'X_i)}
\]

where \( R(t_j) \) denote the risk set of patients at time \( t_j \).

2.4. Accelerated Failure Time Model (AFTM)

Accelerated Failure Time Model[19] assumes a linear relationship between logarithm of survival time and covariates, is given by

\[
\log(T_i) = \mu + \beta'X + \sigma \omega
\]

where \( \beta' \) is a vector of regression coefficients, \( \mu \) and \( \sigma \) are intercept and scale parameters respectively and the error term \( \omega \) is assumed to have some distribution(i.e extreme value, normal or logistic). This transformation leads to the Weibull, Lognormal or log logistic AFT models for \( T_i \).[20]

Now, the survivor function of \( T_i \) is given by
The AFT models are fitted by using method of maximum likelihood. The likelihood of n observed survival times \( t_1, t_2, ..., t_n \) is given by

\[
L(\beta, \mu, \sigma) = \prod_{i=1}^{n} \{ f(t_i) \delta_i \} \cdot \{ S(t_i) \}^{1-\delta_i}
\]

where \( f(t_i) \) and \( S(t_i) \) are the density and survival function for ith patients at time \( t_i \), and \( \delta_i \) is the event indicator function, such that

\[
\delta_i = \begin{cases} 
1 & \text{if the event is observed} \\
0 & \text{if it is censored}
\end{cases}
\]

Results obtained from AFT models can be summarized in the exponentiated form as time ratio (i.e. TR(=exp(\( \beta \))) unlike Cox model hazard ratio. Thus TR>1, associated with prolonged survival time and TR<1, associated with a decrease in survival time.

2.5. Model Comparison

In order to compare parametric AFTM and semi parametric Cox PH Model, we have used Akaike Information Criterion (AIC). The AIC which is a measure of goodness of fit of an estimated statistical model is given by

\[
AIC = -2 \times \log(likelihood) + 2(p+k)
\]

where \( p \) is the number of covariates in the model, \( k=1 \) for exponential and \( k=2 \) for weibull and lognormal models. Model with smaller AIC regarded as a better model.

3. Data Sources

A retrospective follow-up study was conducted, involving 1259 HIV/AIDS patients who were undergoing Antiretroviral Therapy in the ART centre of Dr. Ram Manohar Lohia Hospital, New Delhi, India, during the period of April 2004 to November 2009, and were followed up through the ART routine register records till December 2010. Taking preliminary inclusion criteria as patients should be above 18 years of age and on the basis of availability of complete baseline information on CD4 cell count, date of visit, mode of transmission, weight and hemoglobin etc, 1259 patients were found eligible for the analysis. At the end of the study period, 198 patients were dead and remaining 1061 patients were known to be alive, since the exact date of death is not known, the event death is known to lie in an interval that is the last available CD4 count date and end of the study. Thus the observed event leads to interval censored survival data.

Table 1: presents the descriptive statistics of the study, out of 1259 patients 408 (32.4%) were female. The mean age at diagnosis was 34.24(±8.74) years and the most common mode of transmission was sexual (hetro+ homo) 64.0%, and for 331(26.3%) patients mode of transmission were unknown. 858(68.1%) patients had CD4 cell count less than 200 cell/mm³ at the time of enrollment. 198 (15.7%) patients died and remaining were alive at the end of the study.

The Figure 1: shows NPMLE, K-M survival plot for left, right and mid-point imputation. Comparing these curves we can see that the K-M survival plot for mid-point imputation and NPMLE for intervals are very similar. Then assuming these data as right censored at mid-point of each interval, Cox Proportional Hazard Model and AFTM have been employed to study the effect of prognostic factors on the survival time.

Table 2: presents the results of Cox PH model. The prognostic factors viz. sex, CD4 cell count, past smokers, baseline hemoglobin and baseline weight were found to be statistically significant (P< 0.000). Patients who had CD4 cell count less than 200 mg/mm³ (HR=3.31, 95% CI: 2.19, 4.98), sexually infected (HR=1.12, 95% CI: 0.75, 1.65) and current smokers (HR=2.14, 95% CI: 1.37, 3.34) had an increased hazard for death.
Table 2. Cox-PH Model for HIV/AIDS patients

<table>
<thead>
<tr>
<th>Variables</th>
<th>$\beta$</th>
<th>HR</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.012</td>
<td>0.987</td>
<td>(0.97, 1.02)</td>
<td>0.130</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.780</td>
<td>0.458</td>
<td>(0.32, 0.64)</td>
<td>0.000**</td>
</tr>
<tr>
<td>CD4</td>
<td>1.197</td>
<td>3.311</td>
<td>(2.19, 4.98)</td>
<td>0.000**</td>
</tr>
<tr>
<td>MOT Sexual</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood+IDU</td>
<td>0.113</td>
<td>1.120</td>
<td>(0.75, 1.65)</td>
<td>0.570</td>
</tr>
<tr>
<td>Unknown</td>
<td>-0.185</td>
<td>0.830</td>
<td>(0.60, 1.14)</td>
<td>0.258</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>0.760</td>
<td>2.140</td>
<td>(1.37, 3.34)</td>
<td>0.000***</td>
</tr>
<tr>
<td>Past</td>
<td>0.177</td>
<td>1.193</td>
<td>(0.75, 1.89)</td>
<td>0.451</td>
</tr>
<tr>
<td>Never</td>
<td>-0.201</td>
<td>0.817</td>
<td>(0.76, 0.87)</td>
<td>0.000***</td>
</tr>
<tr>
<td>HB</td>
<td>-0.066</td>
<td>0.935</td>
<td>(0.92, 0.96)</td>
<td>0.000***</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
<td>3628.829</td>
<td></td>
</tr>
</tbody>
</table>

*indicates significance at the 5% level, ** at 1% level and *** at 0.1% level

Figure 2. Cumulative hazard plot of Cox-Snell residual for AFT models.
The parametric AFTM results have been presented in table 3, besides the factors found significant in Cox model, unknown mode of transmission was also found to be a significant predictor in Weibull and lognormal AFT model. Patients who had CD4 cell count <200 cells/mm³ have shorter survival time than patients with CD4 cell count more than 200 cells/mm³ (as TR<1), female patients had longer survival than their male counterparts (TR>1). An increase in weight (in kg) leads to survival than patients with CD4 cell count more than 200 cells/mm³. The mortality is inversely proportional to CD4 count, cumulative probability of AIDS and death increased substantially with decreasing CD4 cell count[23, 28, 29].

Another important result of our study is that patients with sexual (hetero or homo) mode of transmission had worst survival than patients with blood and intravenous drug user mode of transmission, which is departed from earlier result of[21], where they found that intravenous drug user mode of infection had worst survival. However, Remafedi et al[27] have shown that there were no significant differences between deceased and other subjects in relation to mode of transmission. Unknown mode of transmission was found to be a significant factor in weibull AFTM and Lognormal AFTM result.

Patients who are current smokers had association with more hazards of immune deterioration. We found that

### Table 3. Parametric Accelerated Failure Time Model (AFTM) for HIV/AIDS patients

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Exponential</th>
<th>Weibull</th>
<th>Lognormal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>TR</td>
<td>95% CI</td>
</tr>
<tr>
<td>Age</td>
<td>0.012</td>
<td>1.012 (0.99, 1.04)</td>
<td>0.008</td>
</tr>
<tr>
<td>Sex</td>
<td>0.734</td>
<td>2.083 (1.47, 2.94)**</td>
<td>0.534</td>
</tr>
<tr>
<td>CD4</td>
<td>-1.220</td>
<td>0.295 (0.19, 0.44)**</td>
<td>-0.820</td>
</tr>
<tr>
<td>MOT Sexual</td>
<td>-1.159</td>
<td>1.772</td>
<td>1.269</td>
</tr>
<tr>
<td>Blood-IDU</td>
<td>-0.152</td>
<td>0.850 (0.58, 1.27)</td>
<td>-0.086</td>
</tr>
<tr>
<td>Unknown Smoking</td>
<td>0.171</td>
<td>1.186 (0.86, 1.64)</td>
<td>0.239</td>
</tr>
<tr>
<td>Current</td>
<td>-0.769</td>
<td>0.463 (0.29, 0.72)**</td>
<td>-0.562</td>
</tr>
<tr>
<td>Past</td>
<td>-0.138</td>
<td>0.870 (0.54, 1.38)</td>
<td>-0.132</td>
</tr>
<tr>
<td>HB</td>
<td>0.199</td>
<td>1.220 (1.14, 1.30)**</td>
<td>0.144</td>
</tr>
<tr>
<td>Weight</td>
<td>0.062</td>
<td>1.066 (1.04, 1.09)**</td>
<td>0.046</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.159</td>
<td>1.772</td>
<td>1.269</td>
</tr>
<tr>
<td>λ</td>
<td>1.047</td>
<td>1.156</td>
<td>0.909</td>
</tr>
<tr>
<td>γ</td>
<td>0.683</td>
<td>1.464</td>
<td>1.21</td>
</tr>
<tr>
<td>AIC</td>
<td>3359.18</td>
<td>3316.122</td>
<td>3322.259</td>
</tr>
</tbody>
</table>

*indicates significance at the 5% level and ** at 1% level, Time Ratio (TR), TR<1, indicates a decrease in survival time and TR>1 indicates a prolonged survival time.

4. Discussion

In this study we have tried to estimate the survival of HIV/AIDS patients under interval censoring mechanism, since the event of interest death lie in an interval i.e the last available CD4 count date and end of the study. Due to lack of well known statistical methodology and available statistical software, we have adopted imputation approach to handle interval censored data, so that the interval censored data converts to right censored data to which the standard technique can be applied. We have observed that survival function obtained by NPMLE (E-M) for interval censored data has a very close resemblance to the survival function obtained at the mid-point of the interval by Kaplan-Meier method. This is corroborated with the findings of previous studies[10, 17]. However our aim was to determine the effects of prognostic factors age, sex, MOT, baseline CD4 cell, baseline hemoglobin, baseline weight and smoking habits on the survival.

The prognostic factors viz. sex, CD4 cell count, past smokers, baseline hemoglobin and baseline weight are found to be statistically significant (P< 0.000) by both semi parametric Cox PH and parametric AFT Model. Most of the previous studies have suggested that the age is a significant prognostic factor[21, 22, 23]. As age increases the survival time of HIV/AIDS patients decreases. Old age is associated with high risk of disease progression but in our analysis age is not found to be a significant prognostic factor. Also females are observed to have better survival than their male counterparts. As reported previously female had higher life expectancies than male[21, 24, 25, 26]. Remafedi et al[27], had proposed that sex of the patients does not have any significant effect on survival time.

Consistent with the published literature CD4 cell count is found to be an important prognostic marker of HIV/AIDS patients. Patients with CD4 cell <200 cells/mm³ have 3.31 times more hazard to die than patient who had CD4 cell more than 200 cells/mm³. The mortality is inversely proportional to CD4 count, cumulative probability of AIDS and death increased substantially with decreasing CD4 cell count[23, 28, 29].
baseline hemoglobin was a significant predictor of HIV/AIDS patients. To our knowledge hemoglobin has never shown to predict mortality of patients on ART in India, further studies needed to confirm our findings. Baseline hemoglobin level can be used as a simple and practical tool for initial risk assessment in the absence of CD4 cell count and viral load, as was identified in earlier studies by[30] in Tanzania and[31] in Europe. Patient’s weight is positively associated with survival, this is corroborating to the findings of[22] that patients improved clinically with regard to weight and hemoglobin.

There are many situations where AFTM provides better description of data than Cox PH model[32, 33, 34]. Our aim of this study was to compare the performance of both semi-parametric Cox PH model and Parametric AFTM. Based on AIC, Weibull AFT model is found to be efficient (smallest AIC) among the parametric and Cox PH model.

There are some limitations of our study, firstly complete information on patients treatment, follow up data on clinical parameter were not available. Secondly, we have analyzed only one ART centre data, so generalization of our findings would need further conformity.

REFERENCES


