Modeling Cure Rate of Infectious Disease with or Without Co-Infection: An Application to Tuberculosis / Human Immuno Virus

Olaosebikan, A.¹; Aderoju, S.A.²* and Balogun, O.S.³

¹, ²Department of Statistics and Mathematical Sciences, Kwara State University, Malete, P.M.B. 1530, Ilorin, Kwara State, Nigeria. ³Department of Statistics and Operations Research, Modibbo Adamawa University of Technology, Yola, Adamawa State, Nigeria.

ABSTRACT

In this study, we examined the challenges of modeling infectious diseases using tuberculosis (TB) as a case study. The tuberculosis and tuberculosis co-infected with Human Immuno Virus (HIV) is one of the common health problems in the world. Time-to-event outcomes are common data type in medical research. The data examined time until a patient is cured of the disease having some patients right censored. With the nature of the data, the appropriate analysis is survival analysis method. The study aims at fitting appropriate models to the TB and TB/HIV co-infection data examining age and gender as factors influencing the cure rate of the disease. Hence, Kaplan-Meier estimation, Cox PH and some parametric models were adopted in the study. The result shows that among the parametric models, generalized gamma fit TB data best and there is no significant difference in the survival rate of male and female while gamma fit TB co-infected with HIV best and there is a significant difference in the male and female patient. However, Cox PH model (having smaller AIC) performs better than all the parametric models considered (for both data) in this study though with the same conclusion.

Keywords: Survival analysis, TB, HIV co-infection, parametric, Kaplan_Meier, Cox PH
1. Introduction

An infectious disease is a disease that is caused by the invasion of a host by agents whose activities harm the host's tissue and can be transmitted to other individuals. There are six major types of infection agents, they are: bacteria, viruses, fungi, protozoa, helminths and prions\(^1\).

Tuberculosis is a potentially serious infectious disease that mainly affects the lungs\(^2\). The bacteria that cause tuberculosis (TB) are spread from one person to another through tiny droplets released to the air through coughs and sneezes. HIV infection is also a powerful risk factor for TB and contributes to the development of active and latent TB and exogenic re-infection\(^3\). Nigeria is among the 14 high burden countries for TB, TB/HIV and multi-drug resistant TB. It is ranked 7th among the 30 high TB burden countries and second in Africa. The theme for 2019 world's TB day was "It's Time".\(^4\) Pathophysiology and clinical presentation of TB in children differ from what is obtained in adults. Following infectious children have a higher risk not only of progression to disease but also of extra-pulmonary dissemination and death.

Abera\(^5\) worked on the relationship between HIV and TB in Oromia Regional state, Ethiopia. The study covered a total of 40,779 TB cases including 12,818 smear positive pulmonary TB cases and 29,950 positive of HIV infection. The ecological association between different tuberculosis types and prevalence of HIV across zones and towns in Oromia was estimated adopting the Spearman’s correlation. The result of the study showed that the HIV infection prevalence was significantly associated with the incidence of TB in Oromia region \((r = 0.69, p < 0.01)\). It also showed that similar associations were noticed between HIV infection prevalence and the incidence of smear positive tuberculosis and smear negative tuberculosis as well as Extra-Pulmonary Tuberculosis. Kapata \(^6\) showed that there was a need to explore the social determinant of TB and their association with TB/HIV co-infection was important to be addressed in order to have a maximum impact in the control of TB. The study had its limitation due to the fact that it was conducted in an urban setting only. Straetemans \(^7\) carried out a meta-analysis of cohort studies by selecting relevant articles. The purpose was to assess the effect of TB on mortality in people living with HIV. They pooled overall analysis of fifteen studies estimating the effect of tuberculosis on mortality in PLWHIV which show a Hazard ratio \((HR)\) of \(2.6(95\% CI : 1.8,3.6)\) that indicated impact of TB on HIV in co-infection. Zubairu and Musa \(^8\) conducted a study in Nigeria on the case files of patients with HIV/AIDS from January to December 2006 attending Aminu Kano Teaching Hospital. In the study, Chi-square analysis was used to test the significance of association among the categorical variable: All the variables were significantly associated with TB/HIV co-infection and were included in a multiple logistic regression analysis to determine their individual effects. Mashimbye Lawrence \(^9\) considered TB treatment outcomes in adult TB patients receiving treatment at Rixile HIV clinic in Tinswalo hospital in Bushbuckridge, South Africa. The univariate analysis revealed that for the associations of age, sex, Body Mass Index (BMI), education and Antiretroviral (ARV) treatment, only age, sex and ARV treatment were discovered to predict mortality related to TB. Refera Hailu \(^10\) conducted a study on the survival/death status of HIV/TB co-infected patients who were treated of TB at Ambo hospital from September 1\(^{st}\), 2006 to August 31\(^{st}\), 2011. Cox proportional hazard model covariates used significantly influence the survival of PLWHIV co-infected patients are identified. Fatmawati and Hengki \(^11\) proposed an optimal control on the treatment of the transmission of tuberculosis-HIV coinfection model. The optimality system was solved numerically to illustrate the effectiveness of the treatments.
Gesesew et al.,\textsuperscript{12} indicated that one-fifth of Tb/HIV co-infected patients were deceased and social factors seemed to have significant influence. Fatmawati and Hengki presented a mathematical model on the spread of HIV and tuberculosis (TB) co-infection using the resistance of HIV to antiretroviral (ARV) drugs\textsuperscript{13}. The numerical simulations of the optimal control were also performed to illustrate the results obtained.

According to the study carried out by Janida, et al.,\textsuperscript{14} reported that Abuja, Nigeria had one of the highest proportions of TB-HIV co-infection rates in Sub-Saharan Africa and it was also revealed that the outcome of patients had statistically significant higher mortality. Shobowale, et al.,\textsuperscript{15} determined the demographic characteristics of patients with TB and the rate of TB/HIV co-infection from a total of 100 patient’s records retrospectively and analyzed for over 3-month period. They recommended that Tuberculosis still remain a huge public health threat in Nigeria with attendant challenges in diagnosis and treatment and that it is essential that improved system for the accurate diagnosis of Tuberculosis be employed and treatment strategies be improved on and intensified. Yu, et al.,\textsuperscript{16} utilized mixture cure rate model on group survival data and they observed that the estimate of the cure fraction could be quite sensitive to the length of follow up time and the choice of latency distribution (failure time distribution). They also investigated the effects of various parametric distributions such as the lognormal, log-logistic, Weibull and generalized gamma and they concluded that the estimate of the cure fraction was robust with the generalized gamma distribution. It was suggested that the accuracy of the estimate of the cure fraction is affected when the follow up time long with respect to the median survival time and homogeneity of the observations. Chen \textsuperscript{17} in his work concluded that parametric mixture cure rate models possess the flexibility to accommodate varying treatment effects introduced by therapies with different mechanisms of actions.

Elfaki, et al.,\textsuperscript{18} presented a simple modification of estimating for partly-interval cenosed data using the semi-parametric Cox’s proportional hazards regression models of the sub distribution of a two competing risks models. Rodrigues, et al.,\textsuperscript{19} formulated a latent cure rate model with a repair mechanism for a cell exposed to radiation. The latent approach was a flexible alternative to the models presented by Klebanov, et al.,\textsuperscript{20} and Kim, et al.,\textsuperscript{21} which was along the lines of the cure rate model formulated recently by Rodrigues, et al.,\textsuperscript{22}. Li, et al.,\textsuperscript{23} proposed flexible cure rate models in analysing univariate right-censored based on the assumption that the logarithm of survival time follows generalized extreme value (GEV) distribution with spatial and non-linear covariate effects. The model proposed was very flexible but relatively complex.

2. Overview of Survival Methods and their Application in Medical Research.

In everyday of life, we want to know the time it will take for a person to recover from a particular disease, or time until death of an individual infected with a disease.\textsuperscript{24} This type of situation is called survival analysis method.

Generally, survival method is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. The time could be in days, weeks, months or years from the starting point of follow-up of a person until an event happens; on the other hand, time can imply the age of a person when an event occurs. The event means death, disease incidence, relapse from remission, recovery (e.g., cure) or any assigned experience of intrigue that may happen to a person.\textsuperscript{24} In a survival analysis, we for the most part allude to the time variable as survival time, since it gives the time that an individual has "survived" over some follow-up period. We additionally regularly allude to the event as a failure, in light of the fact that the event of interest more often than not is death,
disease incidence, or some other negative individual experience. In any case, survival time might be "time to cure after a treatment procedure," in which case failure is a positive event (as the case is in this study).

Survival methods consider a key analytical issue called censoring. Generally, censoring happens when we have some data about individual survival time, however we don't know the survival time precisely. If for a given patient, the study ends while the patient doesn't get the event, at that point that patient's survival time is viewed as censored. We realize that, for this individual, the survival time is in any event as long as the period that the individual has been followed, however in the event that the individual encounters the event after the study ends, we do not have the foggiest idea about the total survival time. Right-censored data can happen when an individual's actual survival time is greater than the individual's observed survival time. Left-censored data can happen when an individual's actual survival time is less than or equal to that individual's observed survival time.

**NOTE:**

The random variable, \( T = \text{survival time} \) \( (T \geq 0) \)

Survival function, \( S(t) = P(T > t) \)

Hazard function, \( h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \)

\[
\begin{align*}
h(t) &= \frac{f(t)}{S(t)} = -\left[ \frac{dS(t)}{dt} \right] \\
S(t) &= \exp \left[ - \int_0^t h(u)du \right]
\end{align*}
\]

When either of the survival function or hazard function is known, one can get the other.  

### 2.1 The Cox Proportional Hazards (PH) Model

A semi-parametric model is one whose functional form is unspecified. One of the most popular semi-parametric model is the cox proportional hazard model often called Cox PH model which was proposed by David Cox. The cox model formula gives an expression for the hazard at time \( t \) for an individual with a given specification set of explanatory variables represented by \( X \).

\[
h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}
\]

where \( X = (X_1, X_1, X_1, \ldots, X_p) \rightarrow \text{explanatory variables} \)

\( h_0(t) \) is the baseline hazard involving \( t \) but not \( X \).

The PH assumption is that the baseline hazard is a function of \( t \) (time-dependent) but does not involve \( X \)’s. The PH assumption is checked using three methods namely; graphical, goodness of fit and time-dependent variable approaches.

The cox model is widely used because it is "robust" model. Cox is widely popular because it does not rely on distributional assumption for the outcome.

### 2.2 Weibull Model

A parametric survival model is one in which the survival time is assumed to follow a known distribution (e.g Weibull). The time follows a probability density function, \( f(t) \) and the survival and hazard function can be determined. Many parametric models are accelerated failure time (AFT) models rather than PH models.
The Weibull is a two parameter distribution with \( \lambda \) and \( P \) as proposed in 1951 by Weibull.\(^{26}\) It is the most widely used parametric survival model. \( P \) is the shape parameter and determines the shape of the hazard function. The Weibull reduces to exponential when \( P = 1 \). The assumption AFT and PH holds for Weibull.

\[
f(t) = \lambda pt^{p-1}e^{-\lambda t} h_0(t)
\]

\[
S(t) = e^{-\lambda t}
\]

\[
h(t) = \lambda pt^{p-1}
\]

where \( \lambda = \exp(\beta_0 + \beta_1 \text{age} + \beta_0 \text{gender}) \)

### 2.3 The Exponential Model

\[
f(t) = \lambda e^{-\lambda t}
\]

\[
S(t) = e^{-\lambda t}
\]

\[
h(t) = \lambda
\]

### 2.4 Log-normal Model

The log-normal distribution is denoted Log-normal has 2 parameters.\(^{27}\) The shape of the distribution is similar to the log-logistic

\[
f(t) = \frac{\phi\left(\frac{\log(t) - \mu}{\sigma}\right)}{t\sigma}
\]

\[
F(t) = \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)
\]

\[
h(t) = \frac{f(t)}{F(t)}
\]

### 2.5 Log-Logistic Model

The log-logistic is a two parameter distribution \( \lambda \) and \( P \). The log-logistic satisfies the AFT but not the PH model (Kaplan E.L. and Meir P., 1958).

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

\[
h(t) = \frac{\lambda pt^{p-1}}{1 + \lambda t^p}
\]

*recall,

\[
f(t) = h(t)S(t)
\]

### 2.6 Gamma Model

Gamma is a parametric model with two parameters used to model survival data. Gamma is an extension of the factorial function with its argument shifted down by 1 to real and complex number. It can also be used to model service time, lifetime of objects and repair time. The gamma distribution has an exponential right-hand tail. The gamma distribution with parameters \( \lambda \) and \( k \), denoted as \( \Gamma(\lambda, k) \), has probability density function:

\[
f(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)}
\]
\[ S(t) = 1 - I_k(\lambda t), \]

where \( I_k(z) \) is the incomplete gamma function, defined as

\[ I_k(z) = \int_0^z \lambda^{-1} e^{-\lambda t} \frac{dz}{\Gamma(k)} \]

### 2.7 Generalized Gamma Model

This is a parametric survival model with three parameters allowing for great flexibility in its shape. The Weibull and log-normal distribution are special cases of the generalized gamma distribution. The \( S(t) \) and \( h(t) \) are expressed in form of integrals. A log-normal is not a proportional odds model although it satisfies the AFT model. One advantage of parametric model compared to cox model is that the parametric likelihood easily accommodates right, left and interval censored data unlike the cox likelihood that easily handles right censored data but does not directly accommodates left or interval censored data. Binary regression is used for interval censored data and for discrete survival analysis. The density of the distribution can be written as:

\[ f(t) = \frac{\lambda p(\lambda t)^{p-1} e^{-(\lambda t)^p}}{\Gamma(k)}, \quad \text{where } p = \sigma^{-1} \]

The generalized gamma model has the following as its special cases:

- When \( p = 1 \), we have gamma model
- When \( k = 1 \), we have Weibull model
- When \( p = k = 1 \), we have exponential model
- When \( k \to \infty \), we have log-normal model

### 3. Application to Medical Research with examples

The data was collected for a period of 15 years from 2000-2015 from University of Ilorin Teaching Hospital (UITH), Kwara State. The example 1 contains 518 observations of Tuberculosis patients with age (in years) and gender (male=1, female=0) recorded. The time (months) until each patient is cured of TB was also recorded. Similarly, the example 2 contains 133 observations of Tuberculosis co-infected with HIV patients. The analysis was done using “survival” and “flexsurv” packages in R statistical software.

### 3.1 Results of Example 1

There are 518 observations in the TB data. The figure below shows the Kaplan Meir’s curve for gender. It can be seen that there is no significant difference between the survival rates of males and females in the data.

![Figure 1: Kaplan Meir curve of TB data for gender](image-url)
Table 1: PH Assumption Test

<table>
<thead>
<tr>
<th></th>
<th>Rho</th>
<th>Chisq</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.0404</td>
<td>0.767</td>
<td>0.381</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.0159</td>
<td>0.114</td>
<td>0.735</td>
</tr>
<tr>
<td>Global</td>
<td>NA</td>
<td>0.906</td>
<td>0.636</td>
</tr>
</tbody>
</table>

Table 1 shows the PH assumption test. The PH assumption was satisfied from the individual and global tests since the $p$-value > 0.05. In Table 2, the Cox model shows that none of the covariates were statistically significant and the AIC was 4652.09. Several parametric models were fitted to the data, in this case we have Exponential, Weibull, Log-Normal, Log-Logistic, Gompertz, Gamma, Generalized Gamma and Generalized-F. Obviously, the Generalized Gamma best fits this data, having the least AIC. Note that, the p-values of the covariates are enclosed in brackets in Table 2. Moreover, none of the covariates was statistically significant.

Table 2: Semi–parametric and Parametric Estimate of TB data

<table>
<thead>
<tr>
<th></th>
<th>Cox PH</th>
<th>Expo</th>
<th>Weibull</th>
<th>L.Norm</th>
<th>L.Logis</th>
<th>Gomp</th>
<th>Gamma</th>
<th>Gen.G</th>
<th>Gen.F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.00246</td>
<td>0.00197</td>
<td>-0.00235</td>
<td>-0.00305</td>
<td>-0.0031</td>
<td>0.0024</td>
<td>0.0024</td>
<td>-0.0025</td>
<td>-0.0025</td>
</tr>
<tr>
<td></td>
<td>(0.349)</td>
<td>(0.449)</td>
<td>(0.274)</td>
<td>(0.214)</td>
<td>(0.184)</td>
<td>(0.394)</td>
<td>(0.267)</td>
<td>(0.256)</td>
<td>(0.263)</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.0621</td>
<td>-0.0503</td>
<td>0.0558</td>
<td>0.0424</td>
<td>0.0315</td>
<td>-0.0660</td>
<td>-0.0508</td>
<td>0.0461</td>
<td>0.0446</td>
</tr>
<tr>
<td>(male)</td>
<td>(0.518)</td>
<td>(0.599)</td>
<td>(0.480)</td>
<td>(0.646)</td>
<td>(0.724)</td>
<td>(0.491)</td>
<td>(0.526)</td>
<td>(0.578)</td>
<td>(0.594)</td>
</tr>
<tr>
<td>LL</td>
<td>-1857.9</td>
<td>-1844.6</td>
<td>-1856.7</td>
<td>-1855.2</td>
<td>-1854.3</td>
<td>-1841.8</td>
<td>-1840.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>4652.09</td>
<td>3721.74</td>
<td>3697.26</td>
<td>3721.47</td>
<td>3718.40</td>
<td>3716.64</td>
<td>3691.63</td>
<td>3691.38</td>
<td>3693.36</td>
</tr>
</tbody>
</table>

3.2 Results of Example 2

The Kaplan Meir curve for the TB/HIV co-infection data was given in figure 2. The figure shows the survival rate of females is significantly different from males. The PH assumption tested as shown in table 3 was satisfied. This is because the p-values for the individual and global test is greater than $\alpha$. The cox model gives an AIC of 855.367 and both age and gender was significant. Parametric models as in the TB data was also fitted and it shows gamma as the best fit. Both covariates were significant which agrees with the Cox model. Table 4 shows the parametric estimate.
Figure 2: Kaplan Meir curve for the TB/HIV co-infection data

Table 3: PH Assumption test

<table>
<thead>
<tr>
<th></th>
<th>Rho</th>
<th>Chisq</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.1475</td>
<td>2.4287</td>
<td>0.119</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0101</td>
<td>0.0115</td>
<td>0.914</td>
</tr>
<tr>
<td>Global</td>
<td>NA</td>
<td>2.5299</td>
<td>0.282</td>
</tr>
</tbody>
</table>

Table 4: Semi-Parametric and Parametric Estimate of the TB/HIV co-infection data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.0271 (0.002)</td>
<td>0.0181 (0.032)</td>
<td>-0.0193 (0.000)</td>
<td>-0.0151 (0.025)</td>
<td>-0.0130 (0.045)</td>
<td>0.0287 (0.001)</td>
<td>0.0177 (0.002)</td>
<td>-0.0177 (0.003)</td>
<td>-0.0177 (0.003)</td>
</tr>
<tr>
<td>gend (male)</td>
<td>-0.4792 (0.020)</td>
<td>-0.3425 (0.079)</td>
<td>0.3467 (0.006)</td>
<td>0.3505 (0.021)</td>
<td>0.3392 (0.022)</td>
<td>-0.5645 (0.007)</td>
<td>-0.3414 (0.011)</td>
<td>0.3415 (0.011)</td>
<td>0.342 (0.011)</td>
</tr>
<tr>
<td>LL</td>
<td>-454.8</td>
<td>-440.3</td>
<td>-443.3</td>
<td>-443.0</td>
<td>-447.4</td>
<td>-439.8</td>
<td>-439.8</td>
<td>-439.8</td>
<td>-439.8</td>
</tr>
<tr>
<td>DF</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>AIC</td>
<td><strong>855.33</strong></td>
<td>915.59</td>
<td>888.55</td>
<td>894.60</td>
<td>894.08</td>
<td>902.81</td>
<td>887.57</td>
<td>889.57</td>
<td>891.57</td>
</tr>
</tbody>
</table>

The Cox model is expressed as:

\[ \hat{h}(t) = \hat{h}_o(t)e^{0.02713 \text{ age} - 0.4792 \text{ gender}} \]
4. Concluding Remarks

Outcome variable such as time until a particular event occurs is very common in medical research. This type of variable is better handle with survival analysis methods. In this case cure rate of tuberculosis with/without co-infection. The results of the semi-parametric cox proportional hazard model and some selected parametric models were obtained. A key reason why the Cox model is widely popular is that it does not rely on distributional assumption for the outcome and robust. The PH assumption was tested to be true in both examples.

In the Example 1 (TB data), Generalized Gamma model yields the best model. It was equally shown that none of the covariate was statistically significant in determining cure rate of the disease. The Cox model outperforms other models for Example 2 (TB co-infected with HIV). Moreover, both covariates (that is, age and gender) are statistically significant. The survival rates for females with TB/HIV co-infection are higher than that of males across age.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

References


