American Journal of Theoretical and Applied Statistics
2019; 8(4): 125-135
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20190804.11
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Estimating the Age at HIV Infection Retroactively in Limited Resource Settings: *A Case Study of Tanzania*

Theresia Bonifasi Mkenda[1, *], Kaku Sagary Nokoe[1], Samuel Githinji Karoki[2]

[1]Department of Mathematics and Actuarial Science, Catholic University of Eastern Africa, Nairobi, Kenya

[2]Department of Mathematics, School of Science and Technology, United States International University–Africa, Nairobi, Kenya

Email address:

theresiab@yahoo.com (T. B. Mkenda), nokoebiomaths@gmail.com (K. S. Nokoe), skaroki@usiu.ac.ke (S. G. Karoki)

[*]Corresponding author

To cite this article:

Theresia Bonifasi Mkenda, Kaku Sagary Nokoe, Samuel Githinji Karoki. Estimating the Age at HIV Infection Retroactively in Limited Resource Settings: *A Case Study of Tanzania*. *American Journal of Theoretical and Applied Statistics.* Vol. 8, No. 4, 2019, pp. 125-135. doi: 10.11648/j.ajtas.20190804.11

Received: June 25, 2019; **Accepted**: July 18, 2019; **Published**: August 10, 2019

Abstract: Estimation of HIV infection time is a crucial step in HIV/AIDS management as it can help to make informed decisions on the best intervention strategies for controlling new infections, and for taking care of the infected individuals. This study demonstrates three approaches for estimating the age at HIV infection in limited resource settings. Using HIV testing history data collected from a sample of 88 HIV positive women in Kilimanjaro region-Tanzania, we developed a model for estimating the most likely age at which HIV infection occurs for women under reproductive age. The sampled data were collected from typical poor resource settings where access to data is very challenging and the gap between last HIV negative test and first HIV positive test is wide. Formulation of the proposed model involved three steps. Through Modified Midpoint approach, we first determined the midpoint of the age at last negative HIV test and the age at first positive HIV test for each subject. Then, the average time at risk prior to infection, taken over all individuals was subtracted from each midpoint value to obtain the distribution of their estimated age at HIV infection (T). In the second step, survival analysis techniques were used to obtain the Kaplan Meier plots and Nelson Aalen cumulative hazards estimates in which the median age for HIV infection and the most risky age were estimated. The plots of Kaplan Meir survival curves for women with different marital status and levels of education helped to assess whether their age at infection were significantly different. In the third step, we used bootstrap estimation procedures to generate 200 samples of random data and obtain the bootstrap median age at HIV infection and its confidence intervals. The estimated median age at HIV infection from survival analysis approach was 28 years while from bootstrap estimation procedures was 27 years. Likewise, the Nelson Aalen cumulative hazards plot indicated that the most risky age for HIV infection is between 18-40 years while the most risky age from bootstrap estimation was 25 to 27 years. The confidence intervals obtained through bootstrap estimation approach was narrower than that obtained from the survival analysis approach, implying that the bootstrap approach gives more precise estimates. Generally, the study findings provide useful information towards the attainment of the 90-90-90 global HIV/AIDS target as it shows where to allocate more resources and establish more focused interventions for HIV/AIDS management and control.

Keywords: Age at HIV Infection, Modified Midpoint Method, Survival Analysis, Bootstrap Estimation Method

1. Introduction

HIV positive patients may survive with HIV infection for quite a long period before they are diagnosed. These individuals may also show no symptoms and may have a good functional status similar to any other HIV free people. As a result, undiagnosed HIV positive individuals continue to be at high risk of re-infection with different type of HIV and at the same time, transmitting the disease to other people unknowingly. The long interval between actual HIV infection and HIV diagnosis can be associated with poor attitudes towards HIV testing. Most people who go for HIV check-up especially in poor resource settings, have clear reasons for requesting such test like adhering to PMTCT guidelines, travel requirements, marriage requirements or blood donation.

Poor attitude towards voluntary HIV/AIDS counseling and testing increases the number of undiagnosed patients and makes the general management of HIV more complicated especially in low income countries.

One way that can help to manage the HIV/AIDS is to determine the age at which people are more likely to be infected so as to establish proper HIV/AIDS intervention strategies that could inform how and where to allocate the available resources for a better and reliable outputs. Lack of enough information about when people are mostly infected can hinder the decision making related to HIV/AIDS interventions and may increase the problem of HIV data paucity. Knowledge about when the HIV infection occurred to an infected person can also help to understand the general epidemiology of the HIV/AIDS and make proper estimates of the number of people who are at risk. In addition, estimation of HIV infection time may help to track possible sources of infection, monitor the epidemiological aspects of the HIV/AIDS for a specific patient, assess the effectiveness of HIV/AIDS policies and guidelines for managing the disease, and establish the best treatment options for people living with HIV/AIDS.

As we struggle to attain the 90-90-90 global HIV target, African countries especially those in the south of Sahara need to combine their efforts to make ensure that the impacts of HIV/AIDS in their countries are minimized. In Tanzania, the HIV/AIDS still rank as number three killer disease (out of 10) and so more efforts are needed to keep our people safe. The UNAIDS data for 2017 indicates that by the end of 2016 about 1.4 million people were living with HIV in Tanzania and the new infections were almost 55,000 people. Another report from the Operational Plan for HIV prevention in Tanzania Mainland (2016-2018) shows that, only (51%) of adults, and approximately (65%) of children living with HIV were already enrolled on antiretroviral therapy (ART) by end of 2015. Since the current HIV guidelines in Tanzania requires that every diagnosed patient to start treatment immediately there is high probability that the HIV infected people who are not in treatment are also unaware of their HIV status and might have been living with such undiagnosed infection for quite a long time [13]. Therefore, we all need to join efforts in designing ways for managing this epidemic for both infected and uninfected populations to minimize its social, psychological, physical and economic impacts in our country.

1.1. Challenges Associated with the Estimation of HIV Infection Time

The estimation of the infection time for HIV patients have been very challenging since infected individuals normally stay for quite a long period without showing any symptoms though the HIV continue to affect the immunity of a person by destructing his/herCD4 cells slowly. During this period of unknown HIV status, the risk of HIV transmission becomes very high as the viral loads of HIV patients tends to be very high in the initial stage of HIV infection. Another challenge in determination of HIV infection time is that even when the early symptoms like flu, coughing and fever occurs, the undiagnosed patient may rarely associate it with HIV infection but rather with other common diseases like Malaria, Pneumonia or mere allergies. This lack of association is motivated by the fact that the early HIV/AIDS symptoms resembles other simple health problem symptoms and at this stage the level of comfortability of HIV patient is still very good as compared to AIDS patients.

Several researchers in developed countries have attempted to estimate HIV infection time by using HIV testing history data for patients who had conducted multiple tests for HIV infection. [9, 11, 14, 16]. Using such data and statistical techniques, researchers were able to estimate the time when a HIV diagnosed patient was infected or the amount of time that the patient has survived with the virus in his/her body. Lack of enough and appropriate data for HIV testing history makes the process of estimating HIV time more complicated in low-income countries with poor resource settings as compared to developed countries [2]. Some statistical methods that were used to estimate HIV infection time in developed countries may not work properly in low income settings where people tends to go for HIV check only when they are forced by circumstances. Possible methods for estimating the HIV infection time, even in situations where the gap between last negative (LN) HIV test and first positive (FP) HIV test is very wide, are proposed in this study.

1.2. Review of the Methods for Estimating HIV Infection Date

Previous scholars have demonstrated different methods for estimating the HIV infection time through biological approaches and statistical methods [1, 3, 5, 7-8, 10]. For those who used biological approaches they assessed some common biomarkers and tests that can tell whether a person was recently infected or not. Such tests include Tests for recent infections (TRIs), BED HIV-1 Capture enzyme Immunoassays, and Recent Infection Testing Algorithm (RITA). For those who used statistical methods they based much on the assessment of the characteristics of the sampled data in estimating the population parameters. However, the combination of biological approaches and statistical methods has demonstrated by several scholars and seems to bring results that are more meaningful.

Example, some researchers created a biologically motivated time-continuous model of the production of BED-specific IgG data and examined critically the common modeling assumption that seroconversion happens at the midpoint between last negative and first positive HIV test results [10]. To achieve this goal, they collected data from longitudinal cohorts of patients who were tested on regular intervals using a maximum time span between the last HIV negative and first positive test, and assumed that the seroconversion occurred at the midpoint of such interval. Their study findings indicated that the date of seroconversion and by inference, date of HIV infection can be estimated using the midpoint approach for cohorts data but not for data which are collected from public health diagnostic testing as

most of the persons who seek public health services often have a clear reason for that HIV test. The challenge with this approach is on the access to quality data since the collection of such longitudinal data with regular intervals requires enough and expensive resources like skilled personnel, sufficient time and a financial stability.

An extension of the modified back calculation method was proposed by the previous studies to make full use of currently available clinical data in determining the undiagnosed time interval for each HIV positive individual by looking at their CD4 depletion rates and estimating the patient's seroconversion year [5]. These scholars assessed the risk of HIV transmission in the population by estimating the undiagnosed interval of each known infection, construct the HIV incidence curves, and apply the modified back-calculation method to estimate the seroconversion year for each diagnosed patient. Based on the adequacy of CD4 count data, they either estimated patient's pretreatment CD4 depletion rate in a multilevel model or projected one's seroconversion year by referring toseroconverts'CD4 depletion rate. To determine the seroconversion year, the researchers randomly selected a CD4 count within the normal reference range of a healthy adult for 1000 times, using it as a starting point for CD4 depletion, and then calculating the seroconversion year of a particular HIV patient (for 1000 simulations) by using equation (1) below:

$$\text{Seroconversion year} = \text{diagnosis year} + \text{random normal CD4} \times \frac{\text{reference range}_{ij} - a}{\text{monthly CD4 slope} * 12} \quad (1)$$

Where$i$represents thegender (male, female), j is ethnicity (1 Asian, 2 White 3 African & others), a is intercept of an individual's regression line in multilevel model and *CD4 slope* is the adjusted coefficient of the individual's regression line in a multilevel model. In applying that formula, they discarded simulation results that fell outside the lower and upper boundaries. The upper boundary was the year of HIV diagnosis and the lower boundaries were either the last negative HIV testing year, the year when the first possible HIV infection case was noticed in Hong Kong (1980), or the year of attaining 12 year-old, which they consider to be the possible minimum age of being sexually active. However this assumption that '12-year age' is a minimum age at which children engage in sexual activities can be misleading as this age varies across the regions and communities depending on their cultural and social settings. Hence, the results may not be universally acceptable. Their study also demonstrated the possibility of reconstructing HIV epidemic curves from clinical data and illustrating the trends of new infections. However, the success of this method depended much on the quality of the recorded clinical data, which are sometimes rare in limited resource settings.

Another study proposed the use of survival times and binomial models in estimating HIV incidence in United States [3]. To achieve this objective, they used Kaplan Meier estimator and Maximum likelihood function to describe the probability of an individual to be in the recent state of infection as a function of time since seroconversion. They first approximated the entry and exit times, which were the limits of the interval between transitions from recent to non – recent time. Then they determined the seroconversion by using the midpoints of the time between last HIV negative and first HIV positive tests assuming that estimated seroconversion times were uniformly distributed within the seroconversion intervals.

Likewise, a study to estimate the time to HIV-1 Infection from Next Generation Sequence (NGS) diversity was conducted in Sweden by using dataset obtained from 11 untreated HIV-1 patients with known infection dates [8]. The selection of patients who were included in this study required among other criteria, to have a relatively well-defined time of infection or a negative HIV test obtained less than two years before first positive test. The findings of the study indicated that development of next generation sequencing could significantly help to determine the precise time since HIV-1 infection even many years after the infection event. This method seems to be a better option than the commonly used biomarkers or biological approaches like RITA and BED assays that only indicate whether the HIV infection is recent or of a long term. However, its applicability is limited to rare HIV patients with good HIV testing history as most of HIV patients especially in low income countries tend to have a longer than two years interval between last negative and first positive tests.

Previous scholars have also suggested the use of a generalizable method when estimating duration of HIV infections by referring to clinical testing history and HIV test results [7]. With this approach, the results obtained through Fiebig stages and'4thgen'modified staging system methods were assessed and compared with the results from the two-step method for estimating HIV infection time. They used the clinical data on quantitative viral load results and linear mixed effects regression to model the viral load ramp-up dynamics of HIV patients with a random slope and intercept. As there are still no assays, which can help to determine the actual HIV infection time, the researchers opted to use the assay–based reference standard, the Date of Detectable Infection (DDI). Thereafter, they applied a two-step method to calculate infection date estimates based on either the viral load information or the rest of the subjects testing history. In the first step, they determined whether the acute viral load – based estimate could be used to estimate the HIV infection time (i.e if the last day at HIV negative test for a particular patient is known). For the patients whose last negative test data were not available, they moved to a second step in which they estimated the HIV infection duration, by analyzing the remaining HIV testing history. These researchers recommended for new methods for estimating HIV infection time for individual patients to improve the clinical and public health management of newly diagnosed HIV cases.

This study proposes a three- step approach for estimating

the HIV infection time as an extension of the two - step method that were used by previous scholars [1, 7]. The first step was the construction of data by using a modified midpoint method while the second and third steps involved the estimation of HIV infection time and the determination of the most risky age range through survival analysis techniques and bootstrap methods. The modified midpoint method and the survival analysis techniques used a single sample data while the bootstrap estimation involved generation of several random samples derived from the real sampled data collected from the poor resource settings. The third step provides better estimates as the interval of the most risk age group seems to be shorter than the one that we obtained through survival analysis procedures.

2. Methodology

This section presents detailed explanation of the procedures that we used to estimate the age at HIV infection for women under reproductive age. In order to estimate the HIV infection time for each diagnosed patient, we employed three approaches and compared their results. In the first approach, we used the single sample data ($n = 88$) collected from the population of HIV positive women and the modified midpoint method to estimate the most likely age at HIV infection. In this approach, we develop a model that helped us to get a distribution of age at HIV infection and its descriptive statistics. In the second approach, we used non-parametric modeling approach for survival data to estimate the most likely age at HIV infection, using the actual single sample data. We plotted the Kaplan Meier curves and estimated the median age at HIV infection directly from the plots. In addition, the Nelson Aalen Cumulative hazards estimates plot enabled us to obtain the most risky age range for HIV infection. With log rank test, we compared the survivor curves for women with different marital status and levels of education. The 95% confidence intervals for the estimated KM curves and for the median survival times were also constructed.

In the third approach, we used the bootstrap estimation techniques to estimate the age at HIV infection. The use of bootstrap estimation approach was necessary as our data violates the normality assumption and contain some outliers. The bootstrap approach also allowed us to generate many random samples drawn (with replacement) from the actual sampled data, with each observation having the same probability $[P(x_i) = 1/n]$ of being drawn. Using MINITAB software, we obtained the bootstrap median for 200 samples, each with 100 observations. We then derived the bootstrap confidence intervals (95% confidence level) for the median age at HIV infection. We finally compared the estimates from the three approaches for estimating age at HIV infection and assessed their precision.

2.1. Order Statistics

Analysis of order statistics is an important component of statistical inference in cases where the sampled or computer generated data for a particular study are re sorted in ascending order like when dealing with medians, quartiles or sample range. Order-statistics-based inferences can be made in several real life situations like in analyzing the distribution of survival data and in the measurement of financial risks [17]. In this study, we are focusing on the median age at HIV infection, which is an example of the order statistics. A sample data of size n drawn from an infinite population with a random variable$T_1, T_2, T_3 \ldots T_n$can be considered as anordered statistics if we rank this random variable T from smallest to largest value. From the ordered values we may obtained another values Y_1 and Y_n which representsthe smallest and largest values of$T_i$respectively.

That is

$$Y_1 = \min_{\mathrm{i}}(\mathrm{T}_1, \mathrm{T}_2, \ldots\ldots\mathrm{T}_\mathrm{n}) Y_n = \max_{\mathrm{i}}(\mathrm{T}_1, \mathrm{T}_2, \ldots\ldots\mathrm{T}_\mathrm{n}) \quad (2)$$

Generally, the continuous probability density $f(t)$of the r^{th}order statistic for a random sample (Y_r) of size n drawn from an infinite population is given by

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!}\left[\int_{-\infty}^{y_r} f(t)dt\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(t)dt\right]^{n-r} \quad (3)$$

Note that in cases where we have only one sample (i.e.$n = 1$) there will be only one possible order statistic r, hence $r = 1$ and so the probability function f (t) will be$g_r(y_r) = f(y_1)$. Also, for the maximum value of $(r = n)$, the probability density function will reduce to

$$g_n(y_n) = n\left[\int_{-\infty}^{y_n} f(t)dt\right]^{n-1} f(y_n) \quad (4)$$

2.2. Sampling Distribution of the Median

The median is a measure of location that is normally used to make statistical inference regarding a particular population distribution when the distribution is skewed, end values are unknown, or when one requires reduced importance to be attached to outliers [9]. Sampling distribution refers to a probability distribution of sample-based statistic obtained from a large number of samples drawn from a particular population. These sample statistics can be mean, standard deviation, proportion or median. In this study, we are interested with the sampling distribution of the median so we will focus more on it. Suppose we consider a large sample drawn from an infinite population with a probability density function $f(t)$. Apart from giving its smallest and largest orderstatistics Y_1 and Y_n, wemay also find itsmedian. The sample median, which in this study is denoted by $\tilde{x}$will be a number such that approximately half of the observations are below that number and another half is above it. As an ordered statistics, we denote the median value by

$$\tilde{x} = \begin{cases} \frac{1}{2}[Y_m + Y_{m+1}], if n = 2m\ (even) \\ Y_{m+1} if n = 2m + 1 (odd) \end{cases} \quad (5)$$

where m is any positive integer.

This implies that the function will have n independent and identically distributed random variables with density $f(t)$

itha non-zero value at the population median $\tilde{\mu}$ and continuously differentiable in a neighborhood of $\tilde{\mu}$.

We therefore conclude that the median theorem should be used only when there is no possibility of using the central limit theorem to estimate the population mean, since practically the sample mean gives better estimates than the median. Likewise, the sampling distribution of the median may behave differently for the non-normal distributions [6]. In cases where normality assumptions fail, the bootstrap estimation may be the best option for making inferences.

2.3. Estimation of Median Age at HIV Infection Using Bootstrap Approach

Bootstrap approach for statistical estimation of sampling distribution is the best choice in situation where the functional distribution of the parent population is unknown or when the normality assumption cannot be met. Bootstrap techniques can help to obtain an approximate sampling distribution of a statistic, conditional on the observed data. While the most common procedure in statistical inference is to use one sample statistics to estimate population parameters, the bootstrap approach allows the researcher to generate several samples (with reference to the actual sampled data) and use the obtained samples statistics in estimating population parameters. There are several real life situations whose distribution cannot be easily specified in advance, one of them being the distribution of age at HIV infection. HIV infection occurs to different individual from a variety of causes. Children may acquire HIV infection from their mothers during delivery or when breastfeed. Youths and adults may acquire HIV infection during sexual intercourses, blood transfusion or through any other mode of HIV transmission. Hence, a drawn sample of HIV positive patients may have a much-dispersed data regarding age at HIV infection such that the common procedures for making inference may not work.

Making inferences with bootstrap approach involves generating random data from the actual sampled data, determination of bootstrap statistics (like mean and median), estimation of the standard error of bootstrap mean or any other sample statistic θ as well as using repeated sampling method to construct bootstrap confidence intervals [15].

3. Results and Discussion

This section presents the summary of the main findings of this study and their general discussion.

3.1. Descriptive Statistics of the Sampled Data

Data for 114 HIV positive women were collected from nine health facilities that offer PMTCT services in Kilimanjaro region, Tanzania. The main reason for choosing the facilities with PMTCT services is the accessibility to data as we were interested with patients who had tested for HIV at least twice. The women who were under reproductive age had higher chances of testing for HIV several times, as they are required by the HIV guidelines and policies to perform at least three HIV tests from conception to delivery. We found that, only 88women were able to recall their age at last HIV negative test (LN) and age at first HIV positive test results (FP). Since we wanted, first to determine midpoint values for each subject's interval we only used data from these 88 in estimating the age at HIV infection. The current mean age of the sampled women was 36 years with the youngest having 22 years and the oldest, 68 years. Their minimum and maximum age at last negative HIV tests were 17 and 52 years respectively, with a mean of 25 years. The mean age at HIV diagnosis (which we also considered as FP) was 30 years while the minimum and maximum age at HIV diagnosis was 20 and 57 years respectively. The standard deviation and standard error for the mean age at HIV diagnosis were 6.9 and 0.74 respectively. In addition, their median age at HIV diagnosis was 29 years while the sample variance for age at HIV diagnosis was 47.631.

3.2. Estimation of Age at HIV Infection by Using Non-Parametric Modeling of Survival Data

We considered our study as observational, with all interviewed subjects infected by the time, they entered the study but the exact ages at which they were exposed to HIV infection were unknown. We also considered the subjects to be in an open cohort since the size of the risk set could increase or decrease over time unlike the closed cohorts in which the size of risk set always decreases over time [4]. Hence, we defined our outcome variable T (age at HIV infection) as a random variable which represents the difference between the age at HIV confirmation (t) and the average time at risk prior to study entryt_r.

That is,

$$T = t - \frac{1}{2}t_r \tag{6}$$

where,

$t =$ The observed time on study (age at HIV diagnosis)

$\frac{t_r}{2} =$Average timeat risk prior to infection, taken over all individuals, which is unknown but contributes to the true survival time.

$T =$ The most likely age at HIV infection for each diagnosed individual.

The probability plot of the obtained T values obtained through modified midpoint approach shows that the estimated age at HIV infection follows a log logistic distribution with location parameter 3.284 and scale parameter 0.1339. The three-parameter log logistic distribution also found to have a good fit of the data with location, scale and threshold parameters of2.593, 0.2680 and 12.99 respectively. Figures 1 and 2 present the log-logistic and 3-parameter log-logistic probability plots respectively.

We used model (6) to estimate the most likely (median) age at which a woman under reproductive age could be infected using single sample data collected from the field. We let T be the random variable which shows the time to failure (age at which a woman is likely to be infected) and

define the survival function at a time point t as

$$S(t) = P(T > t) \Longrightarrow S(t) = 1 - p(T \leq t) = 1 - F(t) \quad (7)$$

where,

$$F(t) = \int_0^t f(u)du$$

Alternatively, T can be described by a hazard function, which is an instantaneous failure rate at any time point t and denoted by

$$h(t) = \lim_{\Delta t \to 0} P_r\left[\frac{t<T\leq t+\Delta t|T>t}{\Delta t}\right] = \lim_{\Delta t \to 0} P_r\left[\frac{(t<T\leq t+\Delta t)}{\Delta t.S(t)}\right] = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)} \quad (8)$$

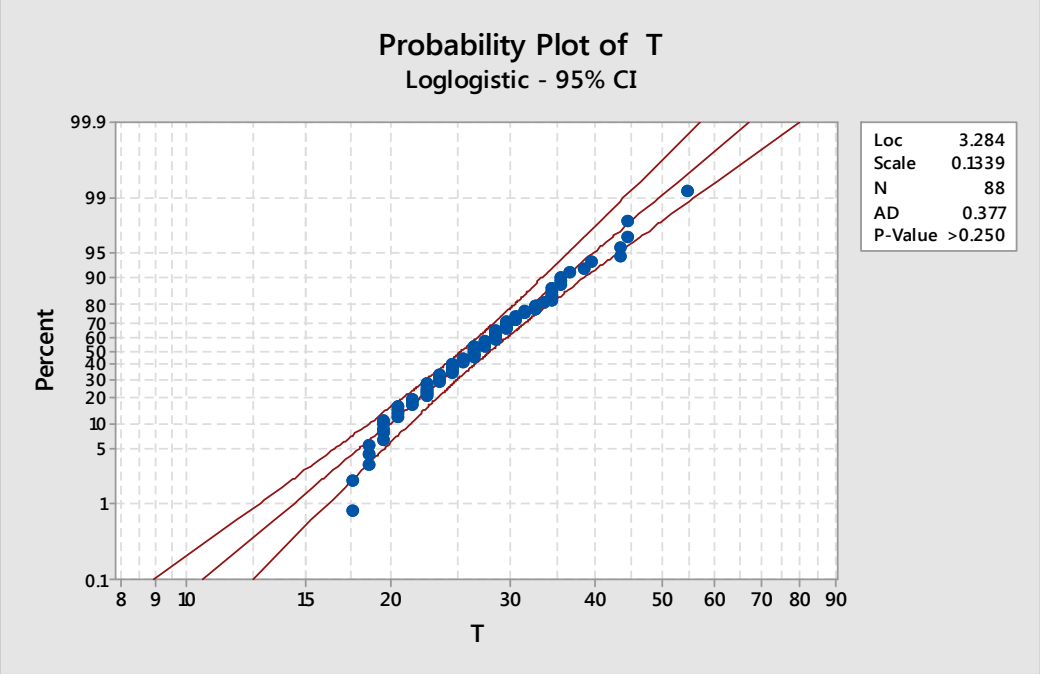

Figure 1. *Log logistic distributional fit for age at HIV infection.*

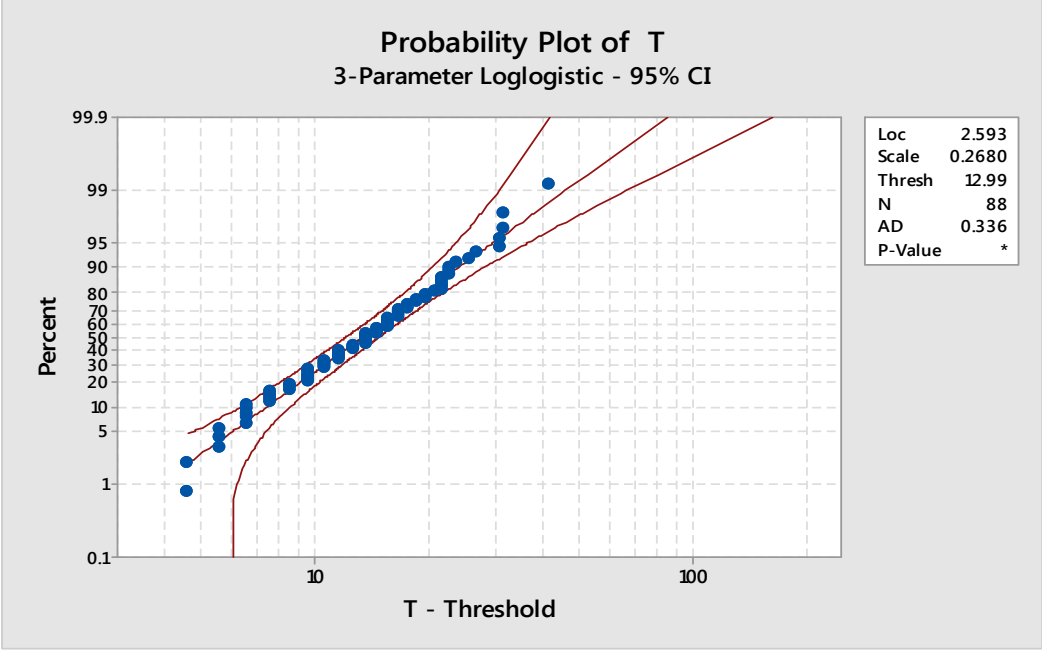

Figure 2. *A 3-parameter log logistic distributional fit for age at HIV infection.*

In addition, survivor function can also be defined in terms of integrated hazard function $H(t)$, derived from the definition of hazard function. That is,

$$h(t) = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} \quad (9)$$

$$\Longrightarrow -\int_0^t h(u)du = \int_0^t \frac{S\prime(u)}{S(u)}du = \log_e S(t) - \log_e 1 \quad (10)$$

$$\Rightarrow -\int_0^t h(u)du = \log_e S(t) \Longrightarrow S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)} \quad (11)$$

The computer layout for our datasets in a simple counting process (CP) format is shown in table 1. The first two columns shows the age at last HIV negative test and age at first HIV positive test respectively. The number of subjects who were at risk of HIV infection at each age point is shown by column three while the remaining columns shows the cumulative failure and hazards for each individual with their respective standard errors and confidence intervals.

Table 1. *Women data layout in a counting process format.*

start	stop	Beg.Total	Cum.Failure	S.e	Hazard	S.e	[95% CI]	
18	19	82	0.0244	0.0170	0.0247	0.0175	0.0000	0.0589
19	20	80	0.0366	0.0207	0.0126	0.0126	0.0000	0.0372
55	56	2	0.9878	0.0121	0.6667	0.6285	0.0000	1.8986
61	62	1	1.0000	.	2.0000	0.0000	2.0000	2.0000

The edited list of Kaplan Meier estimates of the hazard function at the midpoint of each time interval (obtained through STATA package) were presented in a data layout as shown in Table 2. The first column of the table shows the ordered values of the midpoint between age at last negative test and age at first HIV positive test. The second column gives the number of individuals who were still at risk (HIV free) at the beginning of each age interval while the third column shows the frequency counts of those persons who failed at each distinct failure time. None of the subject in the study was lost to follow, so the entries of the Net lost column was zero throughout. The estimates of the hazards function, standard error and the associated confidence intervals for each time interval is shown in column 3, 4 and 5. The data shows that for majority of the subjects in the study, the interval between their last HIV negative test and first HIV positive test was between 0.5 to 2.5 years. Likewise, the STATA output for the confidence intervals for the median survival times under 95% for the midpoint of the intervals found to be in the range 1.5-2.5 years.

Table 2. *Kaplan Meier estimates of the hazard function evaluated at the midpoint of each age interval.*

Time	Beg.Total	Fail	Net Lost	Failure Function	Se	[95% Conf.Int]	
0.5	86	15	0	0.1744	0.0409	0.1090	0.2725
1	71	14	0	0.3372	0.0510	.02480	0.4475
1.5	57	9	0	0.4419	0.0536	0.3445	0.55290
2	48	5	0	0.5000	0.0539	0.4001	0.6095
8	2	1	0	0.9884	0.0116	0.9436	0.9990
9	1	1	0	1.000	-	-	-

Confidence Intervals for the median age at HIV infection

The median of the survival times was calculated by considering the fact that the square of the standardized function of the survival curve around the true but unknown median value (M) is asymptotically distributed in a chi square distribution form. That is, under 95% confidence level,

$$\frac{(\hat{S}_{KM}(M)-0.5)^2}{\widehat{Var}\,[\hat{S}_{KM}(M)]} \sim \chi_1^2 \quad (12)$$

Where, M is the true, but unknown median survival time, $\hat{S}_{KM}(M)$ is the estimated probability from KM curve at the true median survival time and $\widehat{Var}\,[\hat{S}_{KM}(M)$ is the estimated variance of the estimated KM survival probability given by Greenwood 's formula

$$\widehat{Var}\,[\hat{S}_{KM}(M) = (\hat{S}_{KM}(t))^2 \times \sum_{f:t_{(f)}\leq t}\left[\frac{m_f}{n_f(n_f-m_f)}\right] \quad (13)$$

Where t_f is theordered survival times, m_f is number of subjects failing at time t and n_f is the number of subjects in the risk set at the start of the interval. Since we have only one group, the CI for the median will be

$$(\hat{S}_{KM}(M) - 0.5)^2 < 3.84\,\widehat{Var}\,[\hat{S}_{KM}(M)] \quad (14)$$

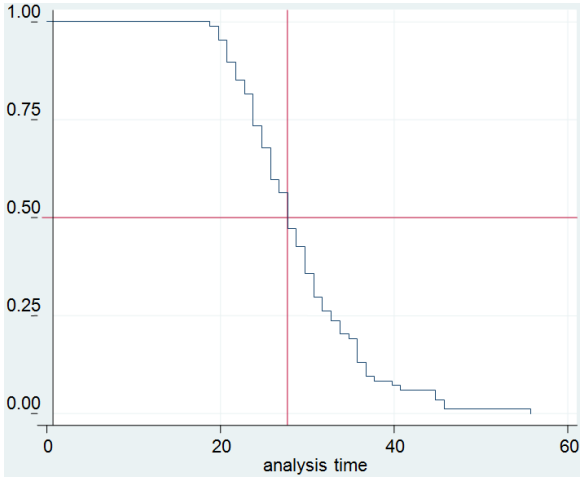

Figure 3. *Kaplan Meier Survivor estimates for age at HIV infection.*

The KM plot for the survival time (estimated age at HIV infection) is shown in figure 3. From the graph, we can determine the estimated median survival time $\hat{S}_{KM}(M)$ as 27.6824, obtained at the intersection of the x-axis value at which y-axis is 0.5.

In addition, the MINITAB plots for cumulative failure,

hazards and survivor functions of the outcome variable T shown in figure 4 indicates that the estimated the median age at which most subjects failed (acquired HIV infection) is 28 years.

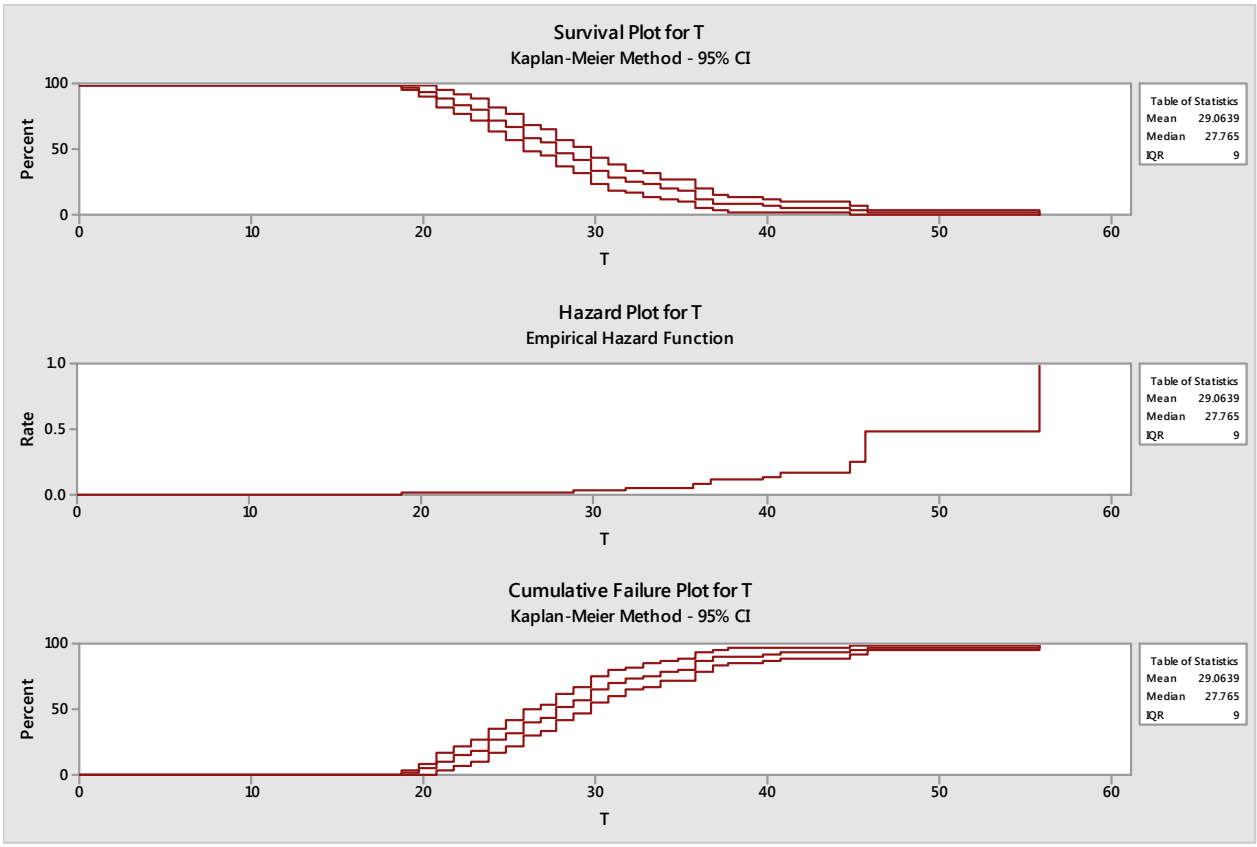

Figure 4. *The survivor, hazard and cumulative plots for age at HIV infection (T).*

In addition to that, the Nelson Aalen Cumulative hazards estimates and the Kaplan Meier failure function plots in figure 5 below indicate a sharp increase in hazards between 18 and 40 years. We therefore consider this age interval as the most risky age group at which most of the women under reproductive age gets HIV infection.

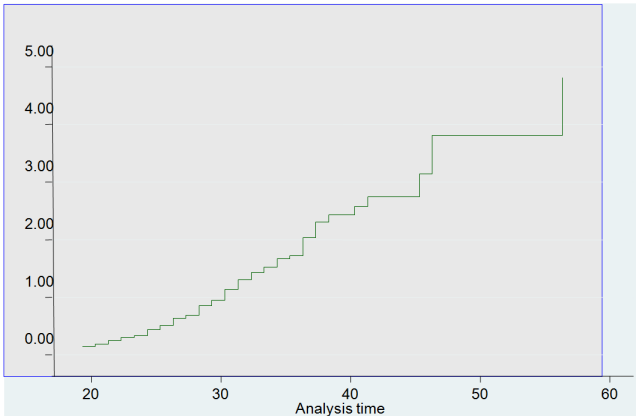

Figure 5. *Nelson Aalen Cumulative hazards and Kaplan Meier plots for women age at HIV infection.*

3.3. Comparison of Age at HIV Infection for Women with Different Marital Status and Levels of Education

We compared the survival curves for women from different groups and assess whether there is a significant difference in their age at HIV infection using the logrank, Wilcoxon, peto and Harringtone tests. The survival curves for women with different marital status are shown in figure 6.

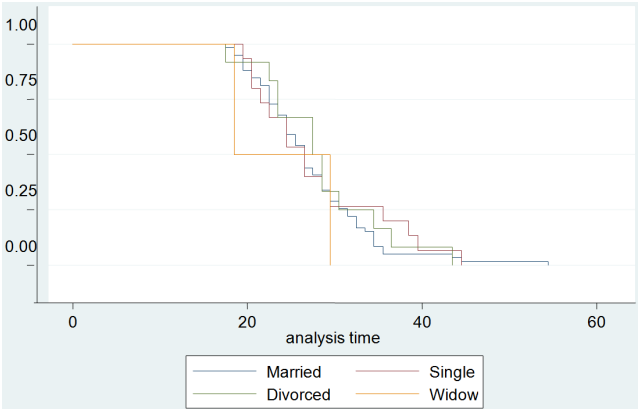

Figure 6. *KM survival estimates for marital status.*

Using the null hypothesis that all survival curves are the same, we performed a log rank test and its various alternative tests for different weightings at 95% confidence level to assess whether the age at HIV infection for women with different marital status were significantly different. These are the variations of the log rank test namely Cox regression based test, Wilcoxon (Breslow), Tarone–Ware, Peto and Flemmington–

Harrington [4]. The use of such tests in comparing the survivor curves for this study was necessary since the age at HIV infection is an example of a clinical problem in which patients are expected to acquire infection at the middle age (especially during the reproductive age range) as compared to young or old ages. This implies the risk of being infected is higher at the later age with the exception of those HIV patients who acquire HIV through vertical transmission. The weights for these different tests are shown in table 3.

Note that the Flemming-Harrington may have different weights for different values of p and q, which are decided by the researcher depending on the nature of the problem. Example if p and q are all zero, then the value of its test statistic will be similar to the log rank test. If p is 1 and q is 0, thenthe test will give more weight to for the earlier survival times where weight will be equals to $\tilde{s}\ (t_{(f-1)})$ andclose to 1. If p is 0 and q is 1, the test will give more weight to for the later survival times where weight will be equals to 1-$\tilde{s}\ (t_{(f-1)})$.

Table 3. *Tests for comparing KM survivor curves.*

S/N	Test Statistic	Weight at f^{th}failure time, $w(t_f)$)
1	Log rank test	1
2	Wilcoxon (Breslow)	n_f
3	Tarone wared cc	$\sqrt{n_f}$
4	Peto-peto	$\tilde{s}\ (t_{(f)})$
5	Flemming-Harrington	$\tilde{S}(t_{(f-1)})^p \times [1-\hat{S}(t_{(f-1)}]^q$

Source: Kleibaum and Klein (2012)

The edited STATA output for the different test results for different marital status and education levels is shown in table 4.

Table 4. *Test results for comparison of survivor curves.*

S/N	comparison test	p-value for marital status	p-value for education levels
1	Log rank test	0.8145	0.7606
2	Cox regression based test	0.8543	0.7928
3	Wilcoxon (Breslow)	0.7973	0.3533
4	Tarone wared cc	0.8116	0.5029
5	Peto-peto	0.7854	0.3330
6	Flemming-Harrington	0.8145	0.7606

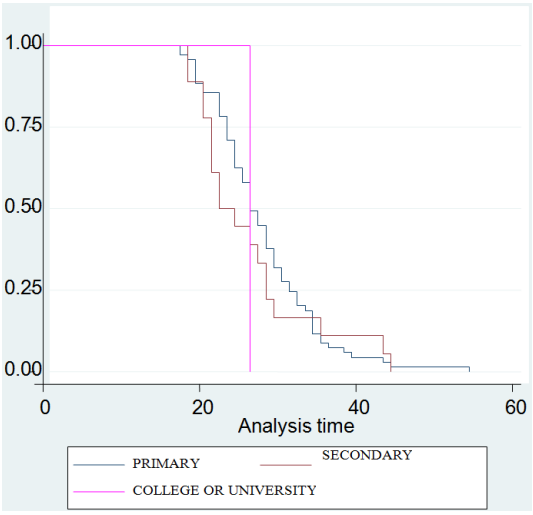

Figure 7. *KM curves for women with different levels of education.*

Since all p-values are greater than the critical value, the test results are not significant enough to make us reject the null hypothesis, so we conclude that the curves for different marital status are not different. Likewise, all of these tests show that there is no significant difference in age at HIV infection for women with different levels of education.

The KM survivor plots for women with different education levels are shown in figure 7.

3.4. Estimation of the Age at HIV Infection by Using Bootstrap Approach

Using MINITAB package, we generated data for 200 samples of size 100 using the original / actual sampled data from the population of HIV positive women in Kilimanjaro. Using bootstrap estimation procedures, we obtained the bootstrap mean and bootstrap median for each of the generated samples.

The edited MINITAB output for the descriptive statistics of the bootstrap mean and median are shown in table 5.

Table 5. *Descriptive statistics for bootstrap mean and median.*

Variable	Total count	Mean (age in years)	Variance	SE Mean	Median (age in years)	Range
Bootstrap mean	200	27.704	0.584	0.540	27.687	4.236
Bootstrap Median	200	26.695	0.641	0.566	26.53	4.000

The summary statistics show that the mean age at HIV infection is approximately 27 years.

3.5. Constructing Confidence Intervals for the Bootstrap Median Age at HIV Infection

To construct the confidence interval for bootstrap sample median we first draw 200 samples from the original sample collected from the field, with replacement by the aid of MINITAB software. Then we computed the sample median for each of the 200 generated samples and ranked the means from smallest to largest value. As finding the 95% implies finding the interval at which the middle 95% of the samples will lie, we had to determine the sample medians at the 2.5% and 97.5% quartiles. The 2.5th percentile was assumed to be at the position 0.025(N+1) while the 97.5th percentile will be at 0.975(N+1). We consider the values at these two positions as the interval at which the bootstrap median age at HIV

infection lies.

So using such procedures we found that 95% confidence interval for the bootstrap sample median were (24.53, 26.53). This implies, women under reproductive age are more likely to be infected with HIV when they are between age 25 and 27 years.

4. Conclusion

This study demonstrates the statistical approach for estimating HIV infection in poor resource settings. This approach involved the use of three different methods namely the modified midpoint method, survival analysis techniques and bootstrap estimation method. Using real datasets collected from nine (9) health facilities located in Kilimanjaro region in Tanzania, we were able to estimate the median age at which women under reproductive age are more likely to be infected by HIV and the most risky age range for HIV infection. Using the modified midpoint method, we found that the average time at risk prior to HIV infection was 2.47 years. The distributional fit of the estimated age at HIV infection (T) found to follow a log logistic distribution with location parameter 3.284 and scale parameter 0.1339. The three-parameter log logistic distribution also found to have a good fit of the data with location, scale and threshold parameters of 2.593, 0.2680 and 12.99 respectively. The KM plot for the estimated age at HIV infection indicated a median survival time $\hat{S}_{KM}(M)$of 27.6824, obtained at the intersection of the x-axis value at which y-axis is 0.5. In contrary, the MINITAB plots for cumulative failure, hazards and survivor functions of T indicated that the estimated the median age at which most subjects failed (acquired HIV infection) is 28 years, which does not differ much with the previous KM plot results. In addition to that, the Nelson Aalen cumulative hazards estimates and the Kaplan Meier failure function plots indicated the most risky age group at which most of the women under reproductive age gets HIV infection is between 18 and 40 years. The comparison of survivor curves for women with different marital status and different educational levels did not show any significant difference in their ages at HIV infection. This implies that we could not find enough evidence to conclude that the probability of a woman infected with HIV under reproductive age is influenced either by her marital status or by education level.

The 95%confidence interval found the most likely age at HIV infection for women under reproductive age was between 18 and 40 years (from survival analysis approach) while for bootstrap estimation was 24.5 and 26.5 years. This implies the bootstrap approach provides best estimates for age at HIV Infection that the survival analysis approach. These study findings provide useful information on where to allocate more resources in the fight against HIV/AIDS epidemic.

Acknowledgements

This study formed part of PhD programme of the main author, which was supported by the Organization for Women in Science in Developing Countries (OWSD). The authors gratefully acknowledge this support.

References

[1] Christopher et al. (2016). A Generalizable Method with an Improved Accuracy for estimation of HIV infection Duration Using Clinical HIV Testing Histories. USA: DOI: 10. 13140/RG. 2. 2. 24528. 30723.

[2] Chinomona, A and Hendry, M (2015): MultipleImputation for Non-response when EstimatingHIV Prevalence Using Survey. BMC Public Health, doi: 10. 1186/s12889-015-2390-1.

[3] Hanson DL, S. R. (2016). Mean Recency Period for Estimation of HIV-1 Incidence with the BED - Capture EIA and Bio-Rad Avidity in Persons Diagnosed in the United States with Subtype B. *PLos ONE*, DOI: 10. 1371/journal. pone, 0152327.

[4] Kleinbaum and Klein (2012). Survival Analysis: A Self Learning Text, third edition. New York: Springer Science + Business Media, LLC. DOI 10. 1007/978-1-4419-6646-9.

[5] Ngai s, e. a. (2016). Estimation of Undiagnosed intervals of HIV Infected Individuals by a modified Back Calculation Method for Reconstructing the Epidem Curves. *Public Library of Science*, DOI: 10. 1371/journal. pone. 0159021.

[6] Omondi, E., mbogo, R., & Luboobi, L. (2018). Mathematical Modelling of the Impact of Testing, Treatment and Control of HIV Transmission in Kenya. *cogent mathematics and Statistics. https://doi.org/10.1080/25742558.2018.1475590*.

[7] Pilcher, et al. (2019). A Generalizable Method for Estimating Duration of HIV Infections Using Clinical Testing History and HIV Testing Results. *AIDS*, DOI: 10. 1097/QAD. 0000000000002190.

[8] Puller V, N. R. (2017). Estimating Time of HIV-1 Infection from Next -Generation Sequence Diversity. *PLoS Comput Biol*, DOI: 10. 1371/JOURNAL. pcbi. 1005775.

[9] Priyanka, K and R, Mittal (2015): Estimation of Population Medianin Two Occasion Rotation Sampling, Journal of Statistics Application and Probability letters, doi. org/10. 12785/jsapl/020304.

[10] Skar H, A. J. (2013). Towards Estimation of HIV-1 Date of Infection: A Time Continuous IgG-Model Shows That Seroconversion Does Not Occur at the Midpoint between Negative and Positive Tests. *PLOS Public Library of Science*, DOI: 10. 1371/JOURNAL. PONE. 0060906.

[11] Stirrup, D. a. (2018). Estimation of Delay to Diagnosis and Incidence in HIV Using Indirect evidences of Infection Dates. *BMC Medical Research Methodology*, https//doi.org/10.1186/s12874-018-0522-x.

[12] Sweeting, M. J, D. Angelis, John P andBarbara S (2014): Estimating the Distribution of the window Period for Recent HIV infections: Acomparison of Statistical Methods, Wiley &Sons Ltd. DOI: 10. 1002/SIM. 3941.

[13] The URT (2013). *National Comprehensive Guidelines for HIV Testing and Counselling.* Dar Es Salaam, Tanzania: National AIDS Control Program (NACP).

[14] Thomas, X. V. (2013). Estimating the time point of acute HCV infection. *Journal of Hepatology*, volume 58, S 206.

[15] Tsokos, R. K. (2009). *Mathematical Statistics with Applications.* United States of America: Elsevier Academic Press Publications.

[16] Wills, K. (2017). Mathematical Modelling Uncovers Mysteriesof HIV Infection in the Brain. *Journal of Neurovirology*.

[17] Zhuo, S (2013): Order Statistics-based Inferences for Censored Lifetime Data and Financial Risk Analysis, PhD Thesis.