AIMS African Institute for Mathematical Sciences CAMEROON

Factors Associated With Mortality In Patients With Ebola Virus Disease

Christabel Lemukong Ngufor (christabel.ngufor@aims-cameroon.org)
African Institute for Mathematical Sciences (AIMS)
Cameroon

Supervised by: Dr. Christiana Kartsonaki, Dr Laura Merson, Dr. Trokon Yeabah
Oxford University, UNITED KINGDOM

29 May 2022

Submitted in Partial Fulfillment of a Structured Masters Degree at AIMS-Cameroon

Abstract

Throughout 2013 to 2016, one of the deadliest Ebola epidemics in record occurred in West Africa, infecting roughly 30,000 people in a year. Aside from the epidemic's extraordinary psychological impact, the epidemic's spillover effect resulted in massive economic losses and multiple deaths. In this study, we identify and investigate the factors associated with mortality in patients with ebola virus disease in West Africa, particularly; Guinea, Liberia and Sierra Leone. This analysis is carried out on a collection of 9472 ebola virus patients from the aforementioned West African Countries. Associations between patients characteristics and other related factors are examined using survival analysis methods and models. Cox regression models, for example, are used to investigate the relation between individual age categories and the duration to death of patients. The results reveal that patients within the age of 41-80 are highly associated with mortality than other age groups. A similar approach is used for the identification of other factors such as sex, country, temperature, respiratory rates which are as well highly associated with mortality. Our findings are suitable for informing policy makers on the ideal measures to take in the face of future epidemics and help drop mortality rates. These findings are equally useful in the development of sensitization campaigns for the different risk groups highlighted.

Keywords: Ebola Virus Disease, Survival Analysis, Mortality, Cox proportional hazard model.

Declaration

I, the undersigned, hereby declare that the work contained in this essay is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Christabel Lemukong Ngufor, 29 May 2022.

Contents

1. Introduction

Ebola virus disease (EVD) is a form of viral hemorrhagic fever (VHF) named after the Ebola River, which was discovered in 1976 in the Democratic Republic of Congo (DRC), in Africa's central region [14]. Among West Africa, EVD is highly transmissible in communities with lack of health-care information, trust, and access. It is a severe multi-system syndrome disease which damages the overall vascular system and is often accompanied with hemorrhage (bleeding) but this is less life threatening compared to the multiple other symptoms where patients usually develop more severe signs such as fever, diarrhea, organ failure, vomiting, muscle discomfort, abdominal pain, and other symptoms during early infection stage and these signs can lead to death between 6 and 16 days of complications. Human-to-human transmission of EVD can occur through direct contact with bodily fluids such as vomit, perspiration, saliva, and breast milk, to mention a few. [17]. EVD is not spread through food and so far, no evidence has shown that mosquitoes or other insects transmit this disease, [13].

The World Health Organization (WHO) labeled this epidemic that started in Gueckedou, Guinea in December 2013, a public health emergency of worldwide concern (PHEIC) on August 8, 2014. The EVD outbreak in West Africa was the largest with over 29,000 recorded confirmed cases in history as of March 25, 2015 and transmission throughout West African countries including Liberia, Guinea, and Sierra Leone [14] which at that time where the most affected countries in West Africa. In Liberia for instance, more than 1.5 million people lived in extreme poverty characterized by poor sanitation, overcrowding, and high crime rates, with an adult literacy rate of less than 43% since over 70% of the urban population lived in slums [10], this is prompt to a high risk of contracting the EVD. These countries lost more than 12% of their gross domestic product (GDP) together with a high fatality rate for Ebola infections, approximately 70% in these areas [12]. In Africa, where the majority of EVD epidemic occurred, many affected persons must travel great distances on terrible roads and in overloaded ambulances to receive EVD treatment. Those sick people who stayed at home due to these problems subsequently contributed to community transmission of EVD due to their extended delay in seeking treatment, while those who made it to the Ebola treatment centers (ETC) died due to their poor health [10].

Previous studies have elaborated on the fact that survival on patients with ebola virus disease depends on long distance travel for search of Ebola treatment centers. Zhang et al [19] suggest that age, chest tightness, paralysis, disorientation, and viral load are all linked to EVD prognosis, with viral load being one of the most important determinants in illness survival. Also, plasma concentrations of EVD are marked different between survivors and nonsurvivors at very early time points after symptom onset and may be predicative of the outcome [12]. Furthermore, Cherif et al [3] found out that patients age was the only independent and significant factor related with mortality of adult patients in Guinea through their analysis. They therefore suggested that older patients should receive special treatment. Nevertheless, all of these results are obtained using statistical tools and most importantly using survival analysis which is a method of analyzing time-to-event data where time-to-event data illustrates the period of time from a well-defined starting point to a well defined end-point of interest [11]. However, the above studies were carried out on a small number of patients and so they had limited variables available on medical records. Therefore, more research is needed to analyze more variables and it would also be necessary to apply survival analysis models to understand multiple variables, which will mean carrying out survival analysis on a bigger dataset.

1.1 Objectives and Contributions

The aim of this study is to identify the factors associated with death of patients due to the EVD, identify patients viral and health care characteristics with risk of death having primary outcome as death and the time of outcome as the time of death or the time of discharge for patients who were discharged from hospital alive. Furthermore, this study will look at the relation between EVD infection and the risk of death during hospitalization using survival analysis models. By training individuals in Liberia, Guinea, and Sierra Leone about the findings of this study, they may be able to assist bring about societal change, and other developing countries how to recognize the factors associated with ebola virus and thereby help improve outcomes. In addition and the lieu of social change, this study will help to inform health authorities in Ebola affected countries what signs and symptoms could indicate a high dead risk, thus informing the triage and prioritization of patient care. It will also serve as a guideline for future attempts to improve EVD outbreak assessments at the household and social levels, as well as health strategies to minimize the rate of EVD transmission and mortality in West Africa.

1.2 Essay Structure

This rest of the essay is organized as follows: In chapter 2, we will have a review of the literature on related articles on Ebola Virus disease. Followed with preliminaries on methods of survival analysis in order to fully understand the statistical results presented in chapter 3. Chapter 3 starts with definition of terms used in survival analysis, describes time to event data using survival functions and the various methods of survival analysis. In addition to this, some tests are presented which is used to compare the various survival curves and conclude based on a certain hypothesis for their significance or differences. Some guidelines on how to build a survival model is also presented in chapter 3. The structure continues with chapter 4 which presents the analysis and results starting with data summary and description of patients characteristics in 4.2. This chapter also involves univariable and multivariable analysis; in general, model building and selection of appropriate covariates associated with mortality and is concluded by some interactions between covariates of different countries and checking for proportional hazard assumptions. The final chapter of this essay is chapter 5 which brings a discussion about the findings from chapter 4 and conclusively section 5.2 which concludes on the results and clearly state out the various factors associated with mortality of patients with Ebola Virus Disease.

2. Literature Review

In the past years, survival analysis tools have been used to study, analyze and understand Ebola Virus disease related issues. Factors linked to the length of stay and treatment result of Ebola patients are included in several of these studies. In this retrospective study conducted in Sierra Leone from September 13, 2014 to November 26, 2014, the the influence of patients duration of symptom time frame, EVD patients' gender, age, occupation, area of residence, and clinical features on the treatment response of 205 laboratory-confirmed EVD patients who were admitted to the Kenema Government Hospital Ebola Treatment Center (KGHETC) were investigated, [10]. The results from this study led to a conclusion that the high length of stay (LOS) for treatment outcome of EVD patients compared to those patients from other districts should be taken in consideration by the health authorities. That is, they should prioritize extensive health education in order to seek early EVD treatment, and they should build strategic Ebola treatment centers as part of their response strategy. However, various constraints, such as a lack of clinical data, made conducting a proper survival study of any relationship between duration of stay and EVD patient referral status difficult, see [10].

Furthermore, in 2017, another survival analysis study was carried out to predict Ebola Severity [8]. The study consisted of analyzing the clinical features of 158 ebola virus positive patients treated at Sierra Leone's GOAL-Mathaska treatment center. Each characteristic's potential was evaluated and combined into a statistically weighted illness score. The mortality rate among the positive EVD patients was 60.8% and highest in those < 5 years old or > 25 years. Moreover, disorientation was the strongest unadjusted predictor of death having odds ratio and p-value of OR = 13.1 and p = 0.014 respectively followed by hiccups, diarrhoea, conjunctivitis and dyspnoea. Including these characteristics in multivariate prognostic scores, Hartley et al. [8] obtained a 91% and 97% ability to discriminate between death at or after triage respectively. Conclusion after analyzing potential interactions predicted a high-risk outcome of death with over 95% accuracy [8]. Additionally, a univariate analysis of predicting mortality of EVD in adults with case study of Guinea identified possible factors associated with mortality outcome to be patient's age which had [OR=1.06, 95% CI=(1.03-1.09)], history of visiting or close contact with a suspected or confirmed ebola virus patient which had p-value of 0.035 and other clinical symptoms on admission such as cough, vomiting, fever, sore throat with odds ratio and 95% confidence intervals [OR=2.49, 95% CI=(0.76-8.14)], [OR=0.97, 95% CI=(0.44-2.16)], [OR=2.32, 95% CI=(0.66-8.16)] and [OR=5.25, 95% CI=(0.96-28.6)] [3], respectively leading to a conclusion that older age was the only independent factor associated with death among EVD adult patients in Guinea. The study's fundamental problem was that all prognostic tools had the potential to become self-fulfilling, and if used wrongly as a medicinal indicator, they had a significant likelihood of overlooking extremely ill individuals, leading to death if the score did not respond to clinical progress.

Mortality in patients with EVD study continues as a Chinese medical team managed EVD patients from Sierra Leone attending to 693 suspected patients, of whom 288 (149 males and 139 females) were confirmed positive of the disease from October 2014 to March 2015 [13]. Clinical symptoms, some manifestations and viral load were analyzed and compared among the different groups for mortality and survival time on the confirmed cases. Among the 288 confirmed patients, the results revealed that the median age was 28 years, and the median log viral load was 6.68 with 98 deaths and 36 recovery; A total of 154 people were lost due to a lack of follow-up. Fever (77.78%), fatigue (64.93%), abdominal discomfort (64.58%), headache (62.85%), and diarrhea (62.85%) were the most common symptoms noticed (61.81%). Patients above the age of 40 had a higher risk of death (odds ratio 2.855, p = 0.044) than those under the age of 18. Patients with a viral load of > 106 copies/milliliters had a greater mortality rate than patients with a viral load of 106 copies/milliliters, with an odds ratio of

3.095 and a p value of 0.004. In addition, Cox regression revealed a relationship between age, viral load, and the prevalence of diarrhea and mortality. Finally, individuals with a high viral load, as well as those who were older and had diarrhea, had a greater fatality rate and a shorter survival duration.

Directing the study in EVD to children. During the 2014-2015 ebola outbreak in Guinea, Cherif et al [2] conducted a retrospective cohort analysis of children under the age of 20 who had laboratory-confirmed ebola virus. 695 cases of EVD were confirmed in the laboratory of the 8448 children with possible or suspected EVD. The CFR (case fatality rate) was 62.9% overall. The death risk was considerably greater in pediatric patients who were younger (OR = 0.995; $95percent$ CI = 0.990-1.000; p = 0.046). Fever (91%), tiredness (87%), and intestinal signs and symptoms (70%) were typical clinical features on admission of pediatric patients, but bleeding signs (24%) did not occur commonly. None of clinical signs and symptoms on admission with epidemiological risk factors for Ebola were associated with mortality outcome which made the study limited. In line with this study, Crowe et al [4] determined 2 readily available indicators which helped to predict survival among young patients with Ebola virus disease in Sierra Leone. Crowe et al, [4] evaluated information for 216 patients of the 227 young patients in Bo District during within a 4-month period. The indicators were time from symptom onset to healthcare facility admission and quantitative real-time cycle threshold (C_t). There were 151 patients alive when they were identified, with documented healthcare institution admissions periods and Ct values. The time from onset of symptoms to admission of healthcare facility was not associated to survival, whereas viral load in the first Ebola virus-positive blood sample was inversely associated with survival. According to the findings, 52 (87%) of 60 patients with a C_t of ≥ 24 survived. These C_t values are very valuable for physicians when making treatment decisions or controlling patient or family concerns.

Abel et [1] conducted a cohort analysis to see if there was a relationship between oral antimalarial medicine and mortality in ebola virus patients. They focused more on the rapid diagnostic test for malaria results and so they were able to do some inferential statistics on the variables from the available dataset and the results were as follows: From patients treated at the five treatment centers, 478 where positive of ebola disease and 424 were eligible for inclusion. The average age was 30.5 years and the proportion of females was 59.7%. Overall, the most commonly observed clinical signs and symptoms were dysphagia 171 (40.3%), dyspnea 135 (31.8%) and diarrhea 298 (70.3%). Furthermore, the findings led them to reach a conclusion based on their objectives, with 376 (88.7%) of the 424 cases received early oral antimalarial treatment. Mortality occurred in $(244)57.5\%$ of all cases. When comparing death rates, early antimalarial treated cases had a 55.1% mortality rate versus 77.1% percent for those who are not treated. Early non-adjusted antimalarial treatment was found to have a lower OR for death than non-treatment with OR = 0.34, $95\%t$ Confidence Interval of (0.12, 0.92) and p = 0.039 . From these results, they arrived at the conclusion that early oral antimalarial treatment in an EVD outbreak was associated with reduced mortality but then further study was proposed to investigate this association between early oral antimalarial treatment and mortality in EVD patients.

All of the aforementioned reviews provide sufficient evidence in knowing the cause of the outbreak of EVD, such as viral load, patient age, and other factors, which enables us to make certain decisions on how to minimize the spread of the disease using survival analysis tools. Nevertheless, we could throw more light in understanding some vital signs of patients with ebola virus disease using survival analysis and methods and models such as semi-parametric and parametric models, Kaplan-Meier estimator and log rank test. The above mentioned methods will help us to better understand how more factors are associated with mortality of patients with Ebola virus disease.

3. Preliminaries

Survival analysis is a kind of statistical analysis that looks at the time leading up to an occurrence rather than the event itself [6]. This time varying is the time of interest, also known as the time of failure or survival. In this chapter, we will present definitions of some terms used in survival analysis, describe time to event situations which will be used practically in the next chapters on a dataset and also state out clearly the different methods and models used in survival analysis.

3.1 Definitions and Examples

3.1.1 Event. Events can be defined as a set of outcomes of an experiment. Another definition is a sample space which indicates all possible outcomes of an experiment. Events are sometimes described as subsets of a sample space.

3.1.2 Random Variable. A random variable is a rule that assigns numerical values to each result in a sample space. There are two types of random variables: discrete and continuous. Only defined values in an interval are assumed by a discrete random variable. Otherwise, the random variable is continuous. Generally, random variables are denoted by capital letters such as X and Y. When X takes values $1, 2, 3, \cdots$, we say that it is a discrete random variable. (see [5]).

A random variable must be measured in the case of a function, which permits probabilities to be given to a set of possible values. Obviously, the outcomes will be influenced by a number of physical elements that are unpredictable. When a fair coin is tossed, for example, the eventual result of heads or tails will be determined by probable physical conditions. Although there are additional possibilities that the coin might break or be lost, such consideration is omitted in this case.

3.2 Censoring

A feature of survival analysis sometimes is that a few subjects might not experience the event under the given observed time frame. Hence, their survival times will not be known to the researcher [16]. There can be some cases wherein the subject experiences a different event which makes it impossible for further follow up. Censoring therefore is simply the act of knowing that the event happened after or before a time interval. Since censored survival data is common, censoring facilitates the application of survival analysis. Censorship takes various forms, each with its own set of events. The following are a few of them:

1. **Right censoring:**

   Areas become right censored in a survival data set when T indicates the time from the onset of the observation to the occurrence of an event. These areas are in a case where the observations break before the event arises. For instance, if T is an event representing the age of a person at death in months. If you know that $T > 60$, the event is appropriately censored at the age of 60. See for example figure 3.1 below.

2. **Left Censoring:**

   Left censoring, on the other hand, is only used when we know T is less than a certain value, say $T < 60$ where 60 is the age of death in years. Left censoring occurs in survival data when some individuals may have already experienced the start time of the observation. Consider a study of the first several days of a girl's menstrual cycle. If you start observing girls at the age of 13 and discover that some of them have already begun menstruation, it means that the cycle's start date is left censored at age 13 only if you can gather information on the start date for those other girls.

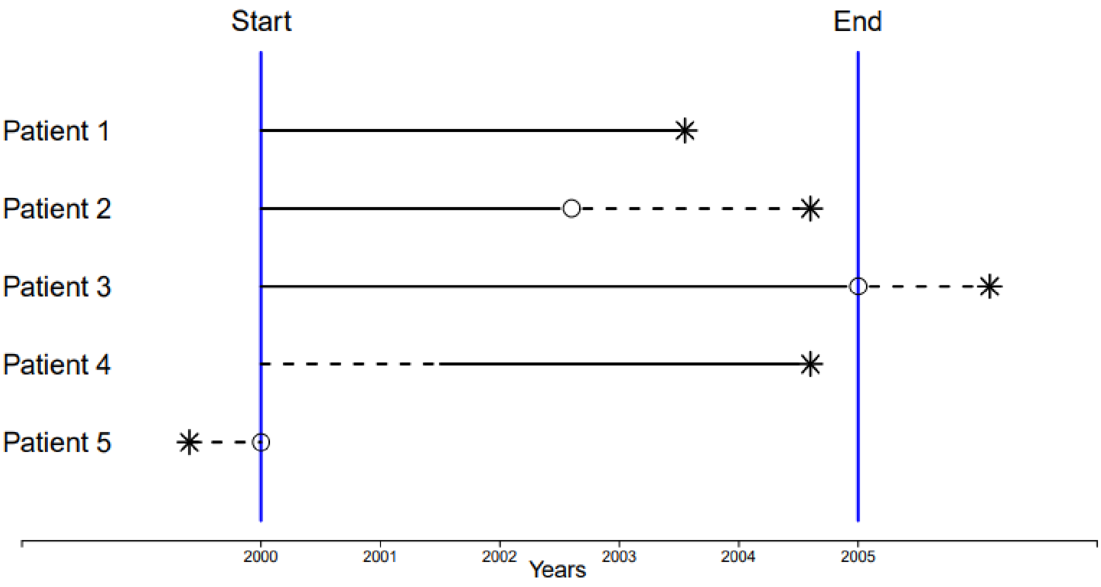

Figure 3.1: Censoring lifetimes of 5 patients in an artificial clinical trial

From figure 3.1, we can observe the following:

(i) The first patient was observed from the beginning of the trial until 3.5 years after the event occurred, hence the time-to-event is 3.5 years.

(ii) After 2.6 years, the second patient moved to a new area, thus we claim Patient 2 had the event after 2.6 years.

(iii) Patient 3 completed the study, thus he or she experienced the incident after 5 years.

(iv) The fourth patient joined the research 1.5 years after it began (late start) then had the event at the age of 4.6 years. As a result, the time to incident is 3.1 years.

(v) Patient 5 experienced the event prior to the beginning of the research. This is an example of left censorship.

3. **Interval Censoring:**

   This censoring reflects a combination of right and left censoring. When you know that the variable T belongs to a specific interval, say $a < T < b$ for some quantities a and b, you can use interval censoring. Once observations are made at a specified time point, this censoring takes place in survival data. Let us consider as an example a case of HIV infection where samples of people are being followed. The time of infection between 8 and 9 would be interval censoring if a person who is not infected at the end of year 8 is then found to be infected at the end of year 9.

3.3 Describing Time to Event

3.3.1 Death Density. Considering a variable to be the length of time taken for an event to occur for example, death, a frequency histogram is constructed which shows the count of events as a function of time [11]. A fitted curve to this histogram gives what we call a **death density function f(t)**. If the area under the curve equal to 1, then the area under the curve to the left of t indicates the proportion of persons in the population who may have witnessed the event of interest at any time, t. The cumulative death function (F(t)) is the proportion of people who died as a function of t.

Particularly, with $T \geq 0$ being a random variable, we have

$$F_\theta(t) = P_\theta(T \leq t) = \int_0^t P_\theta(s)ds \tag{3.3.1}$$

3.3.2 Survival. Considering the death density function, f(t), the area under the curve to the right of time t is the proportion of individuals in the population who have survived to time t. This area is denoted **S(t)** that is the survival time [9]. A survival curve can be created by plotting S(t) as a function of time. The following assumptions are made by S(t):

i. at t=0 , we say there are no failures thus S(t)=1

ii. If there is failure then S(t)=0, and

iii. Note that S(t) is a non increasing function

Let us also note here that survival curves are represented in a step format and the probability of survival is given by

$$S_\theta(t) = P_\theta(T > t) = 1 - F_\theta(t) \tag{3.3.2}$$

3.3.3 Hazard. Conditional failure rate or instantaneous hazard, h(t), is the rate at which an individual chosen at random is observed to be alive at time $(t-1)$ and will die at time t.

Hazard can be defined mathematically as:

$$h(t) = \lim_{dt \to 0} \frac{P_\theta(t < T \leq t + dt | T > t)}{dt} = \frac{P_\theta(t)}{S(t)}, \tag{3.3.3}$$

Where $P_\theta(t) = \dfrac{d}{d\theta} F_\theta(t)$ is a PDF of the random variable T.

That is, hazard describes the 'intensity of death' at the time t given that the individual already survived past the time t. A measure of fatality, death rate, or failure rate is known as instantaneous hazard.

Hazard function can also be written as:

$$h(t) = -\frac{\partial \log S(t)}{\partial t}$$

There is a quantity that is also common in survival analysis, the **cumulative hazard function** which is given by:

$$H(t) = \int_0^t h(s)ds$$

This cumulative hazard can be interpreted as a cumulative amount of hazard up to time t. Cumulative hazard function and survival function are kind of correlated, this link can be written as:

$$H(t) = -\log S(t)$$

And from this it follows that:

$$\begin{aligned} S(t) &= e^{-H(t)} \\ &= e^{-\int_0^t h(s)ds}. \end{aligned}$$

The name hazard is due to some few reasons which are listed below:

(a) It can be easily interpreted.

(b) Can be simplified analytically. That is, it is less complex or complicated.

(c) It models sensitivity, that is it can easily determine uncertainty in the output of a model.

(d) In general, it could be fairly straightforward to understand how the hazard changes with time, for example one can think about the hazard (or death) for a person since his/her birth (see page 8 of [18]).

3.4 Non-Parametric Survival

The first task to perform when the time-to-event data is ready is visualization which allows for pattern recognition of the data and also to identify the appropriate distributional form [7]. In the case of a parametric distribution, the parameters can be determined to explain the survival pattern, allowing statistical inference to be based on the distribution chosen.

3.4.1 Kaplan-Meier Method (KM). Kaplan-Meier method of survival focuses on individual survival times which assumes independence of censoring to survival time. In other words, the reason for censoring the observation does not depend on the failure time.

Kaplan-Meier Estimator is the most popular and simplest non-parametric estimator. The basic computations for the Kaplan-Meier survival curve rely on the computation of conditional survival probabilities.

We start by letting $t_{(1)}, \cdots, t_{(N)}$ be a sorted event time from smallest to largest and assume that S(t) which is the survival time changes only at these sorted times.

In particular,

$$\begin{aligned} P(T \geq t_{(j)}) &= P(T > t_{(j-1)}) \\ &= S(t_{(j-1)}) \end{aligned}$$

So,

$$\begin{aligned} S(t_{(j)}) &= P(T > t_{(j)}) \\ &= P(T > t_{(j)} \quad \text{and} \quad T \geq t_{(j)}) \\ &= P(T \geq t_{(j)}) P(T > t_{(j)} | T \geq t_{(j)}) \\ &= S(t_{(j-1)}) P(T > t_{(j)} | T \geq t_{(j)}) \end{aligned}$$

The estimator is thus given by:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) P(T > t_{(j)} | T \geq t_{(j)})$$

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \hat{P}(T > t_{(j)} | T \geq t_{(j)}), \tag{3.4.1}$$

where each factor

$$\hat{P}(T > t_{(j)} | T \geq t_{(j)}) = \frac{n_i - d_i}{n_i},$$

with n_i being the number at risk, and d_i the number of events at $t_{(j)}$.

3.4.2 Life Table Method. This method is also known as the Cutler-Ederer or Actuarial method. The life table approach is an approximation of the KM method, which is based on grouped survival times. This method assumes that the subjects are randomly withdrawn throughout each interval. Thus, on average we say they are withdrawn halfway through the interval. Also, when we experience a long time interval, it may lead to or introduce bias.

This method also implies that rate of failure within an interval is across all subgroups and is distinct of the likelihood of surviving at other times.

3.4.3 Flemington-Harrington Estimator. Flemington-Harrington estimator explains relationship between cumulative hazard and survival. This estimator says that cumulative hazard function can be used to estimate the survival function. That is;

$$S(t) = exp^{-H(t)}$$

This estimator can be calculated using the **Nelson-Aalen estimator** which is a non-parametric estimator of the cumulative hazard function and is given by [7]:

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_i}{r_i},$$

where the terms can be explained as d_i which is the number who failed out of r_i at risk in interval t_j.

In deriving the Nelson-Aalen estimator, we start by recalling that the Kaplan Meier estimator is written as in equation 3.4.1.

When d_i in equation 3.4.1 is much smaller than n_i, we have the following:

$$e^{-\frac{d_i}{n_i}} \approx 1 - \frac{d_i}{n_i}$$

Therefore,

$$\begin{aligned}\hat{H}(t) &= -\log \hat{S}(t_{(j)}) \\ &= -\log \prod_{t_j \le t} \frac{n_i - d_i}{n_i} \\ &= -\sum_{t_j \le t} \log \frac{n_i - d_i}{n_i} \\ &= -\sum_{t_j \le t} \log e^{-\frac{d_i}{n_i}} \\ &= \sum_{t_j \le t} \frac{d_i}{n_i}\end{aligned}$$

This indicates that the Nelson Aalen estimates the cumulative hazard function by the formula;

$$\hat{H}(t) = \sum_{t_j \le t} \frac{d_i}{r_i} \tag{3.4.2}$$

3.5 Comparing Survival Curves

To provide more helpful information about the differences in time-to-event distributions, understand different categories of covariates and provide an effective process which can help us in identifying such differences. The influential covariates will need to be tested using multivariate analyses.

3.5.1 The Log-rank test. The most popular test for comparing survival distributions is the log-rank test. This test is performed on data that has been subjected to progressive censoring and takes both early and late failures into account. That is, it is assumed that the hazard for both groups is similar.

When a failure occurs, this test calculates the expected number for each event time point. Each table of log-rank test should have observed deaths, expected deaths and calculated variance. The quantities are summed over all the tables to yield a χ^2 statistics with a certain degree of freedom (log-rank test statistic). The observed to expected ratio, which compares the amount of deaths recorded during the follow-up with both the expected number under the hypothesis, is calculated for each group. The null hypothesis states that there is no difference between the survival curves in this test.

Mathematically, Log-rank test is calculated as follows:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Where

$O_1 =$ Overall observed occurrences (example death of patients) in first group

$O_2 =$ Overall oberved events in second group

$E_1 =$ Overall expected occurrences (death) in first group

$E_2 =$ Overall expected occurrences in second group

3.5.2 Other tests. There are two other tests which are sometimes used to compare survival curves or distributions especially when comparing two datasets. These tests are the **Breslow's and Cox Mantel**.

Breslow's test, also known as Gehan's generalized Wilcoxon test, can be used with data that has been progressively censored. When the risk measures are not identical and there is little censoring, it proves in this case to be powerful than the log-rank test. When there is a lot of censoring, its power is poor. It is used when the hazard ratio across groups is not continuous and provides greater emphasis to early failures. The Cox Mantel test, on the other hand, is equivalent to the log-rank test and much more strong than Gehan's generalized Wilcoxon test.

3.6 Non-parametric and Semi-parametric models

The influence of one or more factors on failure time is quantified using survival analysis models. A linear model or the log hazard must be specified in this case. The following parameters can be used to parameterize a parametric model based on the exponential distribution:

$$\log h_i(t) = \alpha + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_k x_{i_k}$$

Equivalently,

$$h_i(t) = exp(\alpha + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_k x_{i_k})$$

The constant α represents the log-baseline hazard in this example, and if $\log h_i(t) = \alpha$, all the $xi's$ are zero.

3.6.1 Cox-proportional hazard model. This is a semi-parametric model where the baseline hazard $\alpha(t)$ is allowed to vary with time no matter if the censored data is discrete or continuous. In this case,

$$h_i(t) = h_0(t) exp(\beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_k x_{i_k})$$

i.e

$$e^{\beta x} = \frac{h_i(t)}{h_0(t)}, \quad \forall t \geq 0 \tag{3.6.1}$$

Equation 3.6.1 is called the hazard ratio which is also known as rate of risk of failure, with $h_0(t)$ being the baseline hazard function which depends on t.

Also, $h_i(t) = h_0(t)$ when all the $x_{i's}$ are zero.

For the set of p-covariates, the generalized proportional hazard model is as follows:

$$h_i(t, x_i) = h_0(t) e^{\beta^T x_i}, \quad t > 0,$$

where

$\beta = (\beta_{1i}, \beta_{2i}, \cdots, \beta_{pi}), i = 1, 2, \cdots, d$ is the regression coefficient.

$x_i = (x_{1i}, x_{2i}, \cdots, x_{pi})^T$ and

$h_0(t)$ the baseline hazard function.

The hazard ratio is used to characterize the outcome of the Cox proportional hazard model. So, the number of times greater (or less) than the hazard of the given reference level of an explanatory variable is the amount of instances greater (or less) than the hazard of an event happening at one level of the explained variable. In other words, this is called a proportional hazards model since the risk for each individual is a fixed percentage of the risk for each other.

Some assumptions of the Cox proportional hazard model are listed below:

i. The ratio of the hazard function for the individuals with different sets of covariates does not depend on time

ii. Time is measured on a continuous scale

iii. Censoring occurs randomly

3.7 Model Building

Model building is a process which is a part of survival analysis which is done in several steps [15].

1. **Selection Of Covariates**

   In this stage, a set of variables is chosen to be included in a survival regression model. Begin by performing a thorough univariate study of the relationship between survival time and each covariate. The survival of groups specified by the variable under consideration is then compared using some significant tests. Furthermore, categorical covariates should be divided into quartiles and the groups should be treated the same way.

2. **Fitting a multivariate model**

   A multivariate model is the one which should contain all covariates which are significant in the univariate analysis with $0.20 \leq p \leq 0.25$ level and any other which seems to be of importance. Following the fit of the multivariate model, the p-value used should be from the **Wald's tests** of the individual coefficients to identify covariates that might be deleted from one model. The **partial ratio test** is then used to confirm if there is significance or not. Also, check if the removal of the covariate is "significant" and the process continues till none of covariates can be removed in the model. A backward process is done again to check if these covariates show evidence of being a confounder or not.

3.8 Testing the proportional hazards assumption

Once a good set of variables has been found, it's a good idea to double-check each one to make sure the cox proportional hazard assumption is correct. We must investigate the extent whereby the predicted risk curves for each layer of stratum of a covariate are distant across time in order to evaluate the proportional hazards assumption.

A graph of a scaled Schoenfeld residuals as a time - dependent and the partial likelihood ratio test can be used to test proportionality of hazards. The Schoenfeld residuals are distributed around 0 in a "well-behaved" model, and a linear regression fitted to the residuals should have a slope of about 0. If the parameters are allowed to fluctuate with time, the fitting curves can be regarded as estimates of S(t).

Alternatively, you can run a separate model for each covariate separately (individually). Then, for that covariate, add a time-dependent interaction term. If the cox proportional hazard assumption holds true for the covariate, then the time-dependent interaction effect will have no effect. The most precise (and objective) strategy for evaluating the proportional risks assumption is to use this method. At this stage, a question arises: what if the variables violate the proportional risks assumptions? In this situation, the other variables can be used to stratify the model, resulting in a different baseline hazard function for each degree of the covariate.

4. Analysis and Results

The statistical program R version $4.1.3$ was used to perform the inferential and descriptive statistics. In this Chapter, we will present some information about the dataset used and the various analysis performed on the data; their results and also interpretations.

4.1 Data Description

The Infectious Disease Data Observatory (IDDO), a scientifically independent, multi-disciplinary researcher-driven network, provided the data for this study. IDDO provides data platforms to capture, share and collaborate on infectious disease related data. It provides the methods, governance and infrastructure to translate data into evidence and improves outcome of patients worldwide. Thus the infectious disease related data used in this study is the Ebola Virus Disease data collected during the 2013-2016 EVD outbreak. The database was made up of a combination of 12 files called domains with patient information which includes demographic information such as gender, age and job; epidemiological history of attending traditional funerals; history of contacting with EVD patients; symptoms and clinical signs.

4.2 Exploratory Data Analysis and Descriptive Statistics

This study consist of records of 9472 ebola virus patients. Table 4.1 shows a table of different percentages of patients characteristic. Results indicates that of the 9472 patients, 45.8% of them were females, 53.7% were males and unknown records were 0.5% of the total patients. The table also indicates that among the 9472 cases, 74.2% (7024) of patients are between the age group of 0-40 years; 24.0% (2270) from 41-80 years; 0.3% (24) are > 80 years with some missing observations. The mean age of the study population is 30.2 years with a SD of 17.1 years. Twenty-nine percent (2751) of the patients where recorded in Liberia; 25.8% (2448) recorded in Guinea and 45.1% (4273) recorded in Sierra Leone. These statistical results can be easily seen from the distributions of patients Age and Country shown in figure 4.1.

The distribution of age is shown in figure 4.1. This distribution is left skewed implying that more patients where recorded from the ages of zero to about 60 years.

4.3 Statistical Analysis

The overall survival of patients is determined using the Kaplan-Meier and Cox regression models (as mentioned in chapter 3). We need to graph the survival curve and analyze how the variables are related to survival in order to determine this total survival. The p value 0.05 is used to determine the significance level. To draw possible conclusions, the risk ratios and 95% confidence level from a model summary of the Cox model will be interpreted. The hazard ratio contains a set of conditions that it must meet: if HR> 1, we have a significant hazard of death (higher risk of mortality); if HR< 1, the risk is low and mortality is low; and if HR=1, there is no influence on the risk of death.

4.3.1 Overall Survival. Overall survival, which is frequently coupled with overall hazard rate, is the likelihood of patients surviving beyond a given time point, let us say t.

Table 4.1: Summary of Patients Characteristics

	Survived (N=6934)	**Died (N=2538)**	**Overall (N=9472)**
SEX			
Female	3144 (45.3%)	1194 (47.0%)	4338 (45.8%)
Male	3747 (54.0%)	1335 (52.6%)	5082(53.7%)
Unknown	43 (0.6%)	9 (0.4%)	52 (0.5%)
AgeGroup			
0-40	5288 (76.3%)	1736 (68.4%)	7024 (74.2%)
41-80	1524 (22.0%)	746 (29.4%)	2270 (24.0%)
> 80	19 (0.3%)	5 (0.2%)	24 (0.3%)
Missing	103 (1.5%)	51 (2.0%)	154 (1.6%)
COUNTRY			
Guinea	1915 (27.6%)	533 (21.0%)	2448 (25.8%)
Liberia	1844 (26.6%)	907 (35.7%)	2751 (29.0%)
Sierra Leone	3175 (45.8%)	1098 (43.3%)	4273 (45.1%)

We can observe from figure 4.2 that the survival curve decreases gradually and suddenly becomes constant around day 20. Also notice that the curve did not reach the median so we cannot say anything about the average survival. This is realistic because about 75% of the 9472 patients survived.

4.3.2 Univariable Survival Analysis. This analysis consists of building models to check the associations of each variable as a response variable with the outcome. We will also include and explain the Kaplan-Meier curves of some variables in this analysis. The results of the univariate models are shown in table 4.2.

Cox proportional hazard regression is the model which is always used to check for associations of both categorical and continuous variables with outcome event. It is also used to model the effect of multiple variables at once. The cox regression model functions both as a linear model and a generalized linear model.

(i) **Age**

This categorical variable was studied in three groups. The division was as follows:

AgeGroup 1:age between 0-40; AgeGroup 2:age between 41-80; AgeGroup 3:age$80+$

Observe from figure 4.3 that the curve of the last age group is not represented. This means that patients above 80 years where few and not recorded within this time of observation. The survival of both age groups from the beginning decreases slowly but after about 10 days, they are both constant. The p-value from this curve indicates the log-rank test value which shows high significance. The Cox proportional hazard as shown in table 4.2 gives the p-values of the different age groups.

In table 4.2 the first age group that is 0-40 is taken as the reference group. Age group 41-80 has coefficient as 0.36342 which indicates this age group has a lower duration than the reference

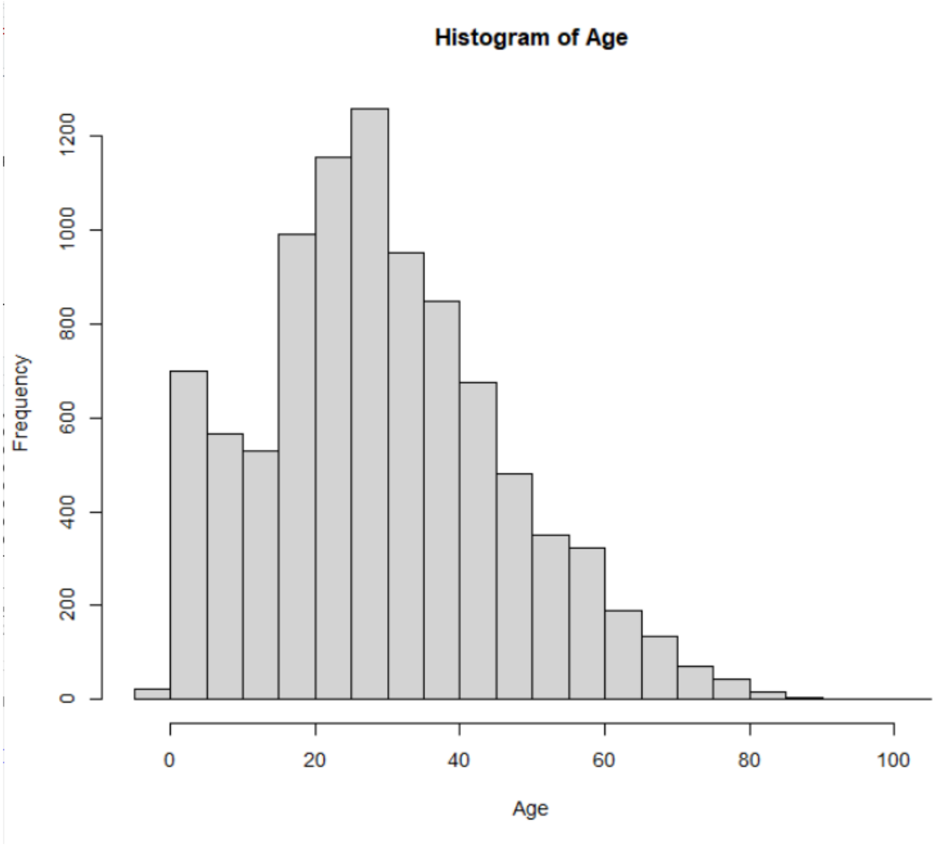

Figure 4.1: Distribution of Age

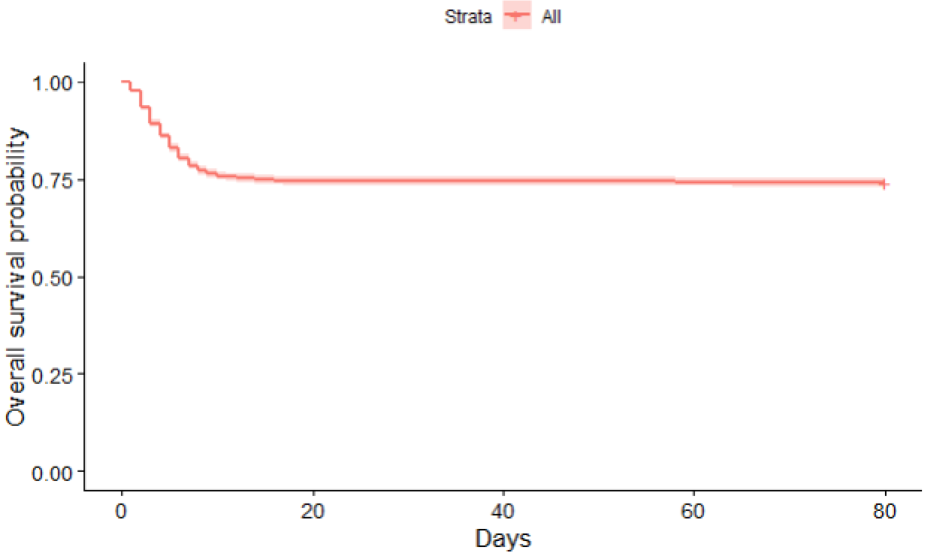

Figure 4.2: Kaplan-Meier plot for overall survival

age and so death will occur faster in patients between 41-80 years than the other age groups. Also from the hazard ratio, we can say that the mortality rate in this group is 1.44 times that of younger and older (above 80) age groups. The p value is very small and thus highly significant.

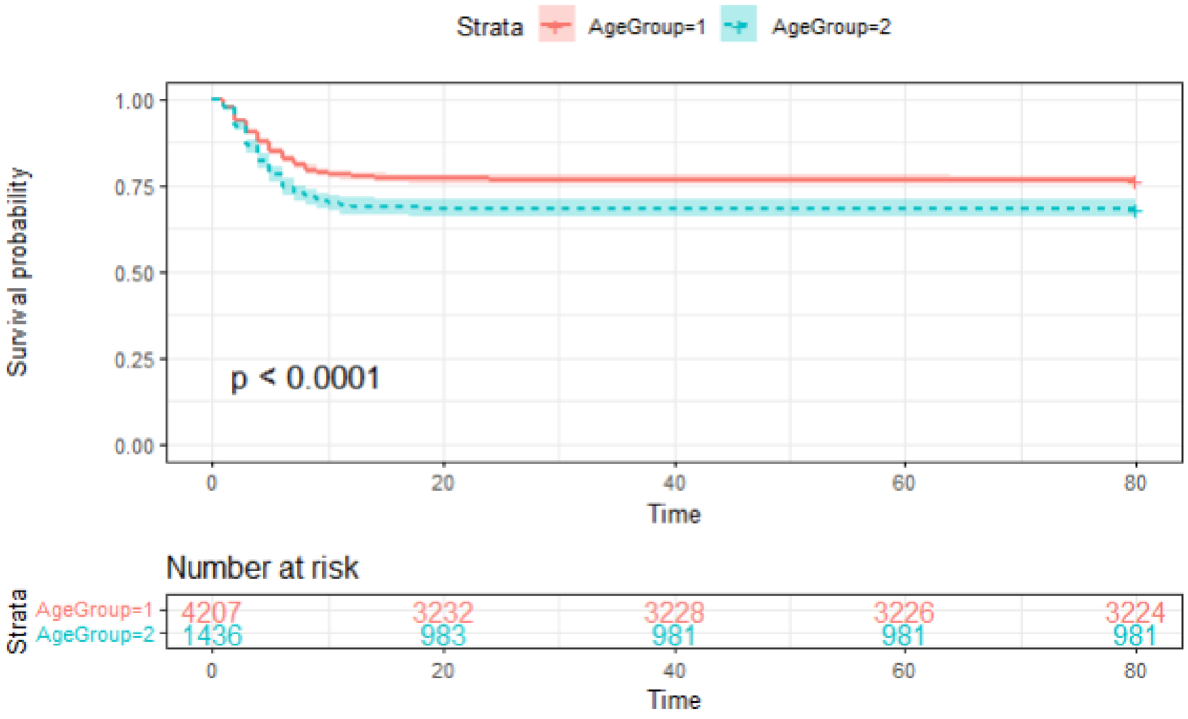

Figure 4.3: Kaplan-Meier curves by different age groups

(ii) **Country**

From the Kaplan-Meier curve in figure 4.4, the survival probabilities are different. Though non of the curves could reach the median line, it is justified since the number of survived patients in this study are more than the deaths. Nevertheless, from table 4.2, taking Guinea as the reference country we can observe that the p values of Liberia and Sierra Leone are significant. Confidence intervals does not include 1 and the hazard ratios tells us that Liberia has a mortality rate of 0.73 times that of Guinea; Sierra Leone has a reduction in the risk of death by 0.40 compared to Guinea. Looking at the parameter estimates of the two countries, they are negative which indicates a high duration of survival in these countries than in Guinea.

(iii) **Sex**

Referring to the Kaplan-Meier curves in figure 4.5, we can notice that the survival curves of both male and female are approximately the same. Table 4.2 takes the female category as a reference with the mortality rate in males to be 0.92 times that of female patients. The 95% confidence interval is (0.8288,1.017) which includes 1 and the p-value is not significant. The coefficient results in this case tells us that the male patients has high duration of survival than the females.

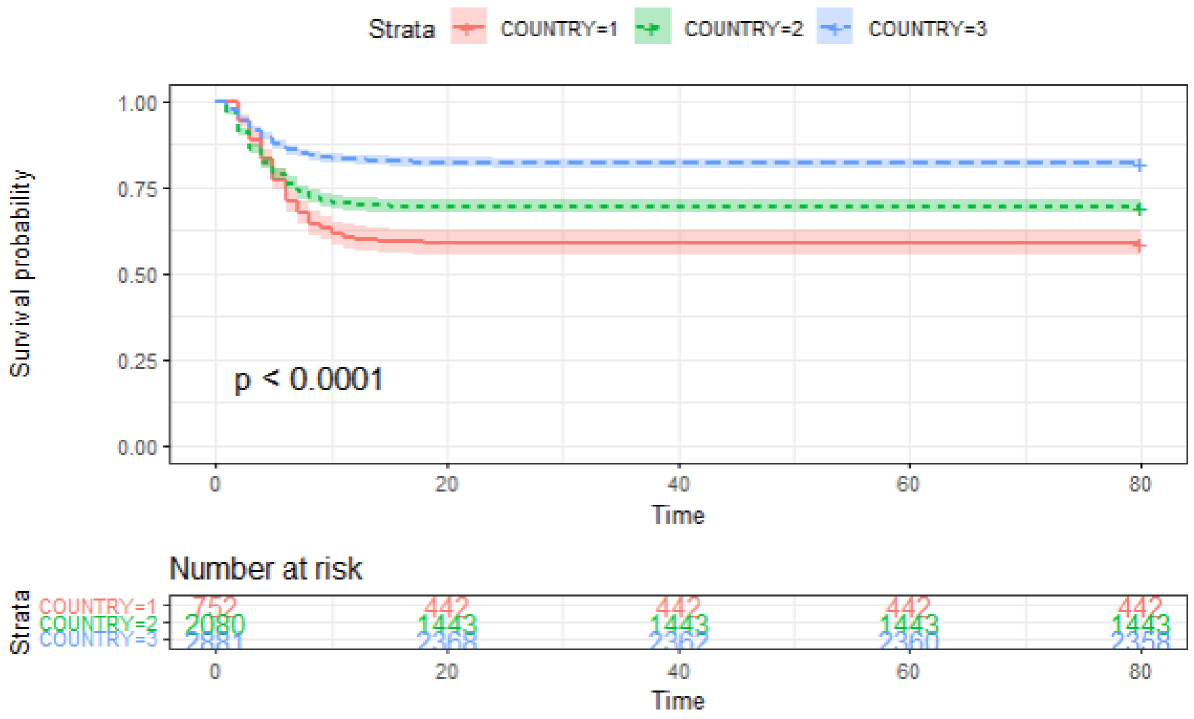

Figure 4.4: Kaplan-Meier curves by country

(iv) **Heart Rate**

Heart rate is divided in to 3 categories:

$$\text{Low} :< 60\text{beats per minute}(bpm); \text{Normal} : 60-100bpm; \text{High} :> 100bpm$$

Referring from table 4.2, patients record were those whose heart rates were high that is more than 100 beats per minute. This model has coefficient of -0.37 indicating that patients with such heart beat have low duration of survival compared to the other heart rates. The p value shows significance and the confidence interval does not contain 1. Also, the hazard ratio is 0.68. Thus patients with such high heart beat have a low risk of death.

(v) **Weight**

Weight is categorized as follows:

$$\text{Group 1} :< 40kg; \text{Group 2} : 40kg - 80kg; \text{Group 3} :> 80kg$$

In our univariate model, under weighted patients are considered as the reference category for weights. From table 4.2, no patients where recorded with normal weight. Overweight patients had coefficient value as -1.0028 indicating that these patients have a high duration of survival implying low risk of death. The p-value indicates significance, confidence interval does not include 1 and the hazard ratio presents rate of risk of death to be 0.37 times that of other weight categories.

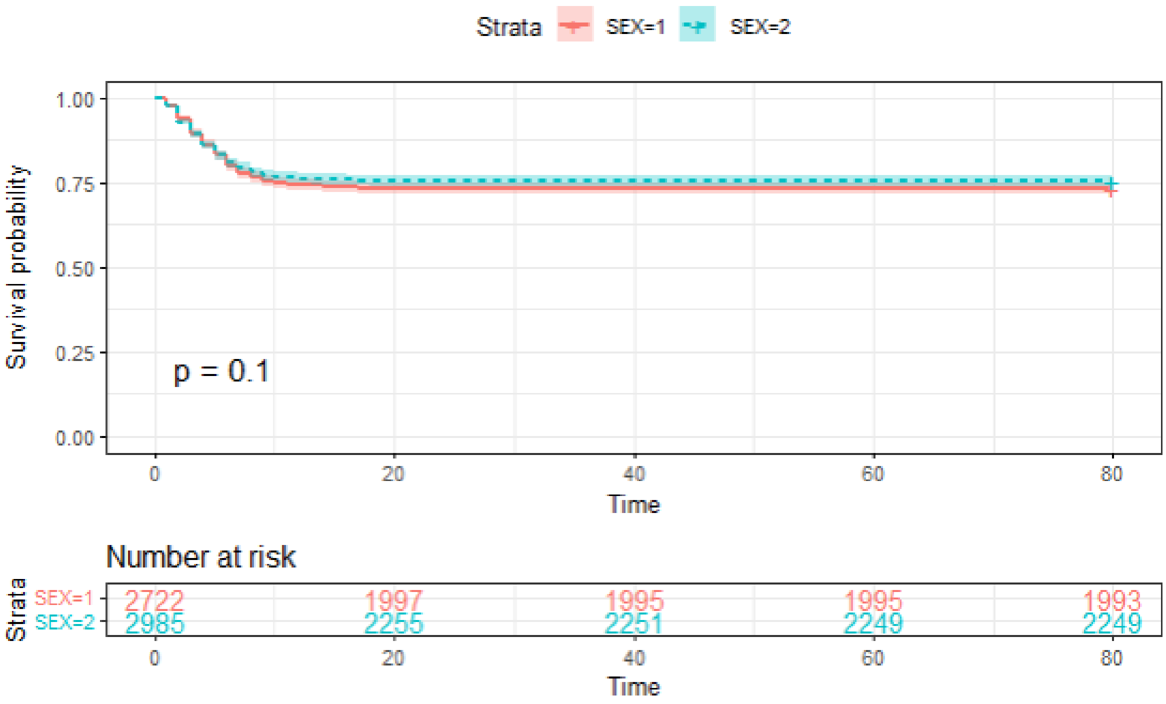

Figure 4.5: Kaplan-Meier curves by sex

(vi) **Height**

Height is categorized as follows:

$$\text{Group 1} :< 1.6m; \text{Group 2} : 1.6m - 1.8m; \text{Group 3} :> 1.8m$$

Cox summary table, table 4.2 gives us the various significance of patients with normal height and tall ones. This covariate presents significance only for tall patients with hazard ratio of 0.24 and specified p-value of < 0.001. Also, 1 is not included in the 95% confidence interval. In this case, we can say that patients with tall heights are associated with mortality in ebola virus disease.

(vii) **Respiratory Rate**

Respiratory rates are measured in breaths per minute and can be categorized as below:

$$\text{Low} :< 12\text{breaths per minute}(bpm); \text{Normal} : 12 - 20bpm; \text{High} :> 20bpm$$

Table 4.2 indicates significance only for patients having high respiratory rates with negative model coefficient thus patients with respiratory rate above 20 breaths per minute have high duration of survival. The confidence interval contains 1 and hazard ratio is 0.77 thus patients with such rate has risk of death as 0.77 times that of other categories.

(vii) **Pulse rate**

Pulse rate is measured in breaths per minute same as respiratory rate but the breaths are taken differently.

$$\text{Low} :< 60bpm; \text{Normal} : 60 - 100bpm; \text{High} :> 100bpm$$

All the different categories of pulse rate seem to be highly associated with mortality since they all have significant p-values, low duration from the model coefficients (positive values), hazard ratios are all greater than 1 and the confidence intervals does not include 1. This is evidence that patients with pulse rate are highly associated with mortality. These results can easily be seen in table 4.2.

(ix) **Temperature**

Temperature of patients was measured in degree Celsius can be categorized as follows:

$$\text{Low} :< 36^oC; \text{Normal} : 36^oC - 38^oC; \text{High} :> 38^oC$$

Table 4.2 presents a result which shows that patients with high temperatures are highly associated with death as the hazard ratio is 1.72 times that of other temperature categories. Also, the model coefficient is positive implying low duration of survival. What the table presents is exactly the same result as can be seen in figure 4.6.

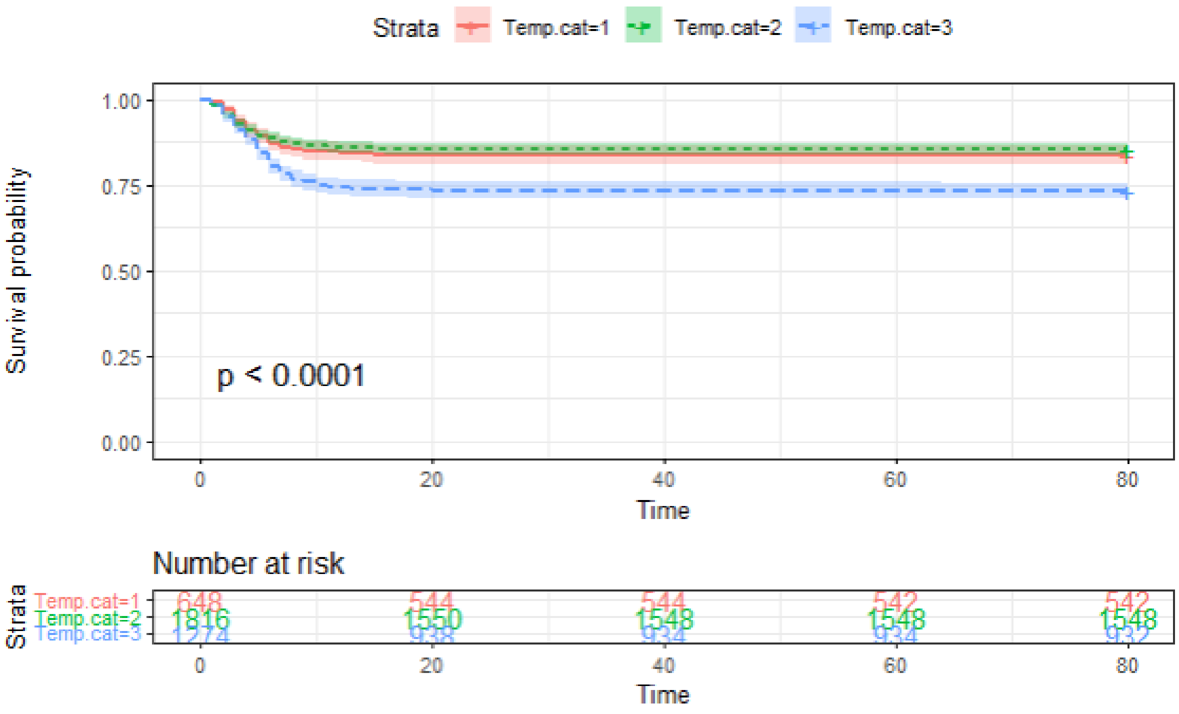

Figure 4.6: Kaplan-Meier curves for temperature category

(x) **Systolic Blood Pressure**

Lastly, we have Systolic and Diastolic blood pressures which were recorded in cmHg thus we categorized as the systolic blood pressure as follows:

$$\text{Group 1} :< 9cmHg; \text{Group 2} : 9cmHg - 14cmHg; \text{Group 3} :> 14cmHg$$

In our case, patients within group 1 and group 2 systolic blood pressures are highly associated with mortality given the results which are presented in table 4.2. That is they have positive model coefficients, significant p values and their confidence intervals does not include 1.

4.3.3 Test Statistics for Comparison. In this section, we'll use a statistical hypothesis test to compare different groups of patients' variables and see if different survival durations are statistically relevant. This test is called the log-rank test statistic as explained in chapter 3, section 3.5. Thus we have the following hypotheses:

H_0 : There is no difference between groups and their survival times.

H_1: There is a statically significant difference between groups and their survival times.

The logrank test is the test statistics which is used to compare survival groups as explained above. The results of this test are presented in table 4.3. This test shows that in the case of oxygen saturation, we reject the alternative hypothesis since the result is not significant and conclude that there is no difference between the groups of oxygen saturation and their survival times. As for the other groups, their p-values are highly significant; in this case therefore, we fail to accept the null hypothesis and conclude that there is a statistically significant difference between the groups of Sex, heart rate, Country, age, temperature, pulse rate, respiratory rate, systolic blood pressure, height and weight with their survival times. The chi-square values of all te covariates are also presented in this table with their corresponding degrees of freedom. The results of sex, age group, country and temperature is exactly what is represented on the kaplan meier curves.

4.3.4 Multivariable Analysis. Multivariable models for this analysis where adjusted for age, sex and country. The results are presented in table 4.4.

The results on table 4.4 are not presented for all the covariates since some adjustments were made. Nevertheless, the same groups of covariates is considered here as for the univariate analysis. Starting with temperature categories, the results are the same with what we have in the univariate analysis. That is, only patients with high temperatures were highly associated with mortality having a hazard of 1.72, p-value < 0.001 and a confidence interval which does not include 1. The results of the different pulse rates of patients are not really clear in this case so we cannot draw any conclusion based on those results. Following with the heart rate, the results show no association with mortality. Thus adjusting for age, sex and country has altered the results of patients heart rate comparing with what we had for the univariate analysis.

In addition, other covariates such as systolic blood pressures, height and weight gives the same results as the univariate analysis. Thus high and low systolic blood pressures together with tall patients and patients with overweight are highly associated with mortality.

4.3.5 Testing for proportional hazard assumption on multivariate models. An important assumption for an appropriate use of the Cox regression model, is to check for violation of the proportionality of hazards. Proportional hazard is specifically to vary over time implying that the effect of a risk factor is constant over time. Several approaches can be used to assess or test this proportionality assumption. In our case we will use graphical assessments.

Let us take a few examples which can be observed from figures 4.8, 4.7, and 4.9. The proportional hazard assumption from these figures is violated. That is, for temperatures and respiratory rates which means that the cox model did not show appropriate results for these covariates. The other plot varies or is proportional over time since the regression line is very close to the zero line hence did not violate the assumption. This violation can be resolved by stratifying the variables that is, building a new model considering those covariates as strata (will not be included in the model).

4.3.6 Assessing associations of covariates separately with different countries. In this subsection, three new datasets where created subset by the different countries. The results are presented in the

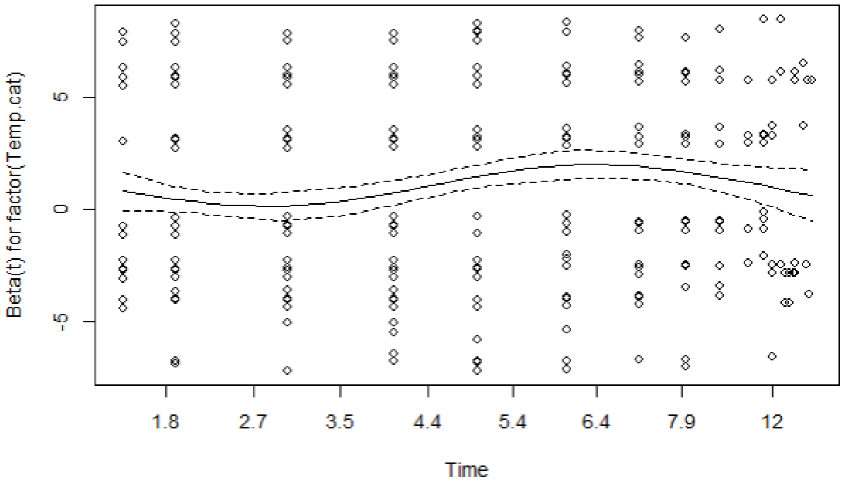

Figure 4.7: Schoenfeld residual plot for temperature

tables 4.5, 4.6 and 4.7.

The results of the model with subset Guinea is presented in table 4.5. We can observe similar results obtained from the univariate analysis.

The male patients in Guinea had less risk of death compared to the female patients. This is seen from the HR = 0.91 and the 95% confidence interval which includes 1. The p-value in this case does not show significance. For the different age groups, patients from Guinea with age group between 41-80 years has a hazard ratio greater than 1 (HR=1.66) indicating that they have a high risk of dying from ebola virus thus highly associated with mortality. We can also notice in this case that the confidence interval does not include 1 and the model coefficient is positive (low survival duration).

In the case of temperature, the risk of death of patients in Guinea is high for those with high temperatures with HR=1.105; the results from pulse rate is a bit unpredictable for patients from Guinea. Across heart rate values to respiratory rate and systolic blood pressures, only patients with high respiratory rate having p-value of 0.04, HR of 1.53 had a significant association with death of patients from Guinea. The rest of variables which has not been mentioned do not have an y association with mortality of patients from Guinea.

The results of the model with subset Liberia are presented in table 4.6. These results are very much similar to that of Guinea. The similarities occur in the male category, age group 41-80 years and high temperature category. Their corresponding hazard ratios brings us to the same conclusion on their high associations with mortality of patients from Guinea.

One more covariate is high systolic blood pressure, where patients with high and normal pressures from Liberia are highly associated with death as the p-value of 0.04 and < 0.001; HR of 1.88 and 2.07 shows high risk of death. The rest of variables which are not mentioned did not have records of patients from Liberia.

The results are referenced in table 4.7. The same observation from patients in Guinea is recorded here but for high heart rate which has p value more than the significant level and the confidence interval includes 1. Results of the respiratory rates are also not significant and high systolic blood pressures

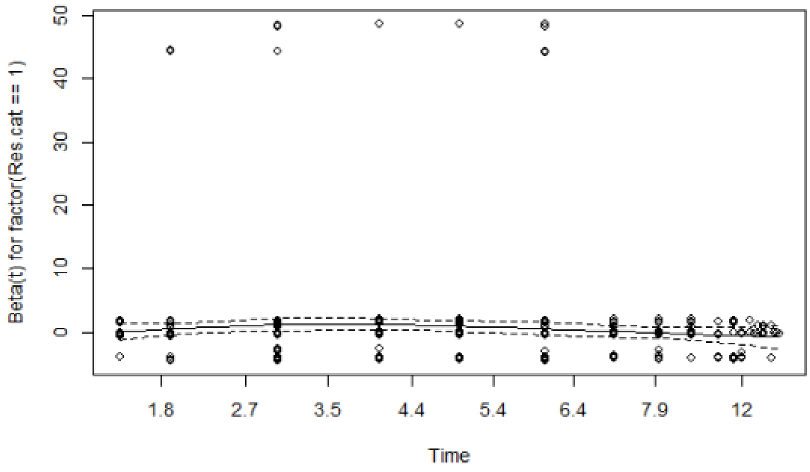

Figure 4.8: Schoenfeld residual plot for respiratory rate

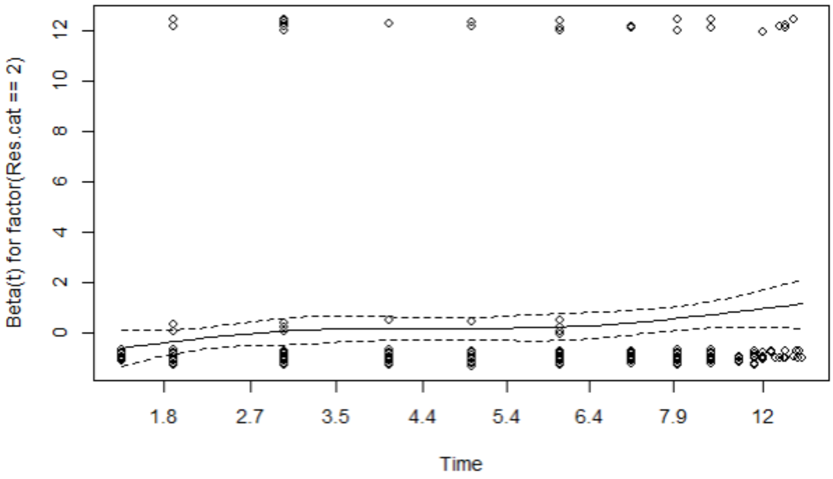

Figure 4.9: Schoenfeld residual plot for respiratory rate

gives the opposite result to what is presented for the other countries. Therefore patients from Sierra Leone having high heart rate, normal and high respiratory rates and high systolic blood pressures are not associated with mortality.

Table 4.2: Univariable Analysis accessing associations of each covariate with the outcome

Variables	Model Coefficient	HR	CI(95%)	P-value
Sex				
Female	ref			
Male	-0.085	0.93	(0.8288,1.017)	0.102
Country				
Guinea	ref			
Liberia	-0.31086	0.7328	(1.282,1.576)	< 0.001
Sierra leone	-0.91543	0.4003	(0.45,0.55)	< 0.001
AgeGroup				
(< 40)	ref			
(41-80)	0.36342	1.438	(1.287,1.607)	< 0.001
($80+$)	–	–	–	–
Temperature				
Low ($<$36°C)	ref			
Normal (36°C- 38°C)	-0.1029	0.9022	(0.7205,1.130)	0.37
High ($>$38°C)	0.5408	1.7173	(1.3811,2.135)	< 0.001
Pulse Rate				
Low ($<$60bpm)	0.8952	2.448	(1.684,3.559)	< 0.001
Normal (60-100 bpm)	0.6350	1.8870	(1.506,2.365)	< 0.001
High ($>$100bpm)	1.0764	2.9341	(1.575,5.466)	< 0.001
Heart Rate				
Low ($<$60bpm)	-	-	-	-
Normal (60-100 bpm)	-	-	-	-
High ($>$100bpm)	-0.36669	0.693	(0.6195,0.7752)	< 0.001
Respiratory Rate				
Low ($<$12bpm)	0.61332	1.8466	(1.1433,2.9823)	0.01
Normal (12-20 bpm)	-0.18096	0.8345	(0.6470,1.0763)	0.16
High ($>$20bpm)	-0.26528	0.7670	(0.6814,0.8633)	< 0.001
Oxygen Saturation				
Low ($<$90%)	–1.2	6.14×10^{-6}	(0, inf)	0.98
Normal (90%-100%)	-	-	-	-
High ($>$100%)	-	-	-	-
Systolic BP				
Low ($<$9 cmHg)	1.0587	2.883	(1.7026,4.881)	< 0.001
Normal (9-14 cmHg)	0.3066	1.359	(0.7507,2.459)	0.31
High ($>$14 cmHg)	0.5270	1.694	(1.3278,2.161)	< 0.001
Height				
Group 1 ($<$1.6m)	-	-	-	-
Group 2 (1.6-1.8m)	0.77907	2.17943	(0.9778,4.8602)	0.05
Group 3 ($>$1.8m)	-1.42141	0.24137	(0.2021,0.2883)	< 0.001
Weight				
Group 1 ($<$40 kg)	-	-	-	-
Group 2 (40-80 kg)	-	-	-	-
Group 3 ($>$80 kg)	-1.00288	0.3668	(0.3219,0.4181)	< 0.001

Table 4.3: Log-rank tests for different covariate groups

	Chi-square	**P-values**
Country	84.8 on 2df	<0.001
Sex	5.5 on 1df	0.02
Age group	28 on 1df	<0.001
Temperature	72 on 2df	<0.001
Pulse Rate	16.4 on 3df	<0.001
Heart Rate	4.7 on 1df	0.03
Respiratory Rate	27.1 on 3df	<0.001
Systolic BP	66.2 on 3df	<0.001
Height	173 on 2df	<0.001
Weight	90.3 on 1df	<0.001
Oxygen Saturation	0.4 on 1df	0.5

Table 4.4: Multivariable Analysis adjusted for sex, age and country

Variables	**Model Coefficient**	**HR**	**CI(95%)**	**P-value**
Temperature				
Low (<36°C)	ref			
Normal (36°C- 38°C)	0.07315	1.0759	(0.8526,1.3576)	0.54
High (>38°C)	0.54254	1.7204	(1.3746,2.1531)	< 0.001
Pulse Rate				
Low (<60bpm)	ref			
Normal (60-100 bpm)	0.5352	1.594	(0.2385,12.2284)	0.594
High (>100bpm)	-1.1900	1.763×10^{-6}	(0,Inf)	0.98
Heart Rate				
Low (<60bpm)	-	-	-	-
Normal (60-100 bpm)	-	-	-	-
High (>100bpm)	0.10918	1.1154	(0.9330,1.3334)	0.23
Respiratory Rate				
Low (<12bpm)	ref			
Normal (12-20 bpm)	0.29976	1.3495	(1.0050,1.8121)	0.04
High (>20bpm)	0.21762	1.2431	(1.033,1.4955)	0.02
Systolic BP				
Low (<9 cmHg)	ref			
Normal (9-14 cmHg)	0.03581	1.0365	(0.5478,1.9611)	0.912
High (>14 cmHg)	0.09132	1.0956	(0.7501,1.6002)	< 0.001
Height				
Group 1 (<1.6m)	-	-	-	-
Group 2 (1.6-1.8m)	0.75654	2.1309	(0.9499,4.7802)	0.06
Group 3 (>1.8m)	-1.36807	0.2546	(0.2073,0.3127)	< 0.001
Weight				
Group 1 (<40 kg)	-	-	-	-
Group 2 (40-80 kg)	-	-	-	-
Group 3 (>80 kg)	-1.01070	0.36397	(0.3047,0.4347)	< 0.001

Table 4.5: Associations of covariates subset Guinea

Variables	**Model Coefficient**	**HR**	**CI(95%)**	**P-value**
Sex				
Female	ref			
Male	-0.09553	0.91	(0.6441,1.295)	0.6
AgeGroup				
(< 40)	ref			
(40-80)	0.5089	1.6634	(1.14,2.482)	0.008
($80+$)	–	–	–	–
Temperature				
Low (<36°C)	ref			
Normal (36°C- 38°C)	-0.1390	0.8702	(0.3454,2.192)	0.768
High (>38°C)	0.1002	1.1054	(0.6697,1.824)	0.69
Pule Rate				
Low (<60bpm)	ref			
Normal (60-100 bpm)	0.5331	1.704	(0.2381,12.2)	0.536
High (>100bpm)	-1.400	8.28×10^{-7}	(0,Inf)	0.993
Heart Rate				
Low (<60bpm)	ref	-	-	-
Normal (60-100 bpm)	-	-	-	-
High (>100bpm)	0.001166	1.012	(0.7051,1.452)	0.95
Respiratory Rate				
Low (<12bpm)	ref			
Normal (12-20 bpm)	0.269	1.009	(0.6043,2.836)	0.49
High (>20bpm)	0.425	1.5296	(1.0188,2.297)	0.04
Systolic BP				
Low (<9 cmHg)				
Normal (9-14 cmHg)	-0.21318	0.808	(0.4032,1.619)	0.5
High (>14 cmHg)	0.04809	1.049	(0.7200,1.529)	0.8

Table 4.6: Associations of covariates subset Liberia

Variables	Model Coefficient	HR	CI(95%)	P-value
Sex				
Female	ref			
Male	-0.08809	0.915	(0.8265,1.014)	0.09
AgeGroup				
(< 40)	ref		–	< 0.001
(41-80)	0.357	1.429	(1.293,1.615)	
($80+$)	–	–	–	
Temperature				
Low (<36°C)	ref			
Normal (36°C- 38°C)	0.0106	0.0181	(0.8048,1.2755)	0.91
High (>38°C)	0.59822	1.81889	(1.4609,2.22647)	< 0.001
Respiratory Rate				
Low (<12bpm)	ref			
Normal (12-20 bpm)	0.15303	1.1054	(0.3724,1.5567)	0.3
High (>20bpm)	0.03915	1.0399	(0.8745,1.2367)	0.7
Systolic BP				
Low (<9 cmHg)				
Normal (9-14 cmHg)	0.6323	1.88	(1.0360,3.419)	0.04
High (>14 cmHg)	0.7272	2.0689	(1.5311,2.795)	< 0.001

Table 4.7: Associations of covariates subset Sierra Leone

Variables	Model Coefficient	HR	CI(95%)	P-value
Sex				
Female	ref			
Male	-0.18160	0.8339	(0.6983,0.9959)	0.04
AgeGroup				
(< 40)	ref			
(41-80)	0.46257	1.588	(1.319,1.913)	< 0.001
($80+$)	$-$	$-$	$-$	$-$
Temperature				
Low (<36°C)	ref			
Normal (36°C-38°C)	0.1231	1.131	(0.8796,1.454)	0.3
High (>38°C)	0.5306	1.787	(1.3896,2.298)	< 0.001
Heart Rate				
Low (<60bpm)	ref	-	-	-
Normal (60-100 bpm)	-	-	-	-
High (>100bpm)	0.1629	1.1770	(0.955,1.45)	0.126
Respiratory Rate				
Low (<12bpm)	ref			
Normal (12-20 bpm)	0.2895	1.336	(0.9687,1.842)	0.07
High (>20bpm)	0.1577	1.171	(0.9538,1.438)	0.18
Systolic BP				
Low (<9 cmHg)				
Normal (9-14 cmHg)	2.1414	8.511	(2.1193,34.18)	0.002
High (>14 cmHg)	0.66388	1.894	(0.2663,13.47)	0.5

5. Discussion and Conclusion

Ebola virus disease was one of the most deadliest disease outbreak in west Africa from 2013-2016 which comes with very minute signs and symptoms some of which are presented in this study to check their associations with death. Death of patients with ebola virus varies depending on different age groups, sex and other factors as discovered in this study. Identifying such factors is crucial but once understood will be able to reduce the rate of mortality of patients with ebola virus disease.

Therefore, we investigated some variables which are associated with the death outcomes among patients using methods of survival analysis. The time of interest was recorded in days of survival of ebola virus patients and the events of interest was death. Comparison was done between distribution of survival times in different groups to examine how much the factors and their groups are different in survival times. Kaplan Meier curves were plotted for overall survival showing how patients who survived after about 20 days upon admission since their survival probability did not go down to zero or even below the average survival time.

Log-rank test was used to test for significance between different covariate groups and it was found that sex, heart rate, country, age, temperature, respiratory rate, systolic blood pressure, weight and height are statistically significant whereas oxygen saturation did not show any significance.

The investigation followed with univariate and multivariate Cox models in which if the sign of the coefficients in Cox model is negative, it will mean that the risk of death is lower while if it is positive then the risk of death is higher. The cox model therefore resulted to a conclusion that female patients, patients with high temperature, pulse rate, high heart rate, low and high respiratory rate, low and high systolic blood pressures, heights $> 1.8m$ and weights $> 80kg$ are associated with mortality of patients with ebola virus disease. This result is similar to what was obtained from the multivariate models but for the fact that in the multivariate model, low systolic blood pressures together with over weighted patients where not associated with mortality. In addition, checking the proportional hazard assumption of some covariates, we notice that there was violation with the temperature model and respiratory rates.

A very interesting finding from the analysis and results is that, associations between these factors including the vital signs with mortality might be different depending on the country in which the patient is from. Results from such findings indicated that there is a slight difference in associations of systolic blood pressures of patients from different countries. This tells us that systolic pressures in general cannot be used to draw conclusions as a vital sign which is highly associated with death of patients which ebola virus disease from West Africa. It should be well noted that our findings presented a parametric model which can be used to fit a distribution and give better estimates to which one can qualify the associations of clinical factors with mortality. This model is the Weibull distribution model.

5.1 Limitations and Recommendation

Throughout this study, we analyzed associations between variables (age, sex, country, vital signs) and the response variable being the outcome. The dataset did not have clinical characteristics thus the study was limited to some medical records and history of patients. Many records such as pregnancy status, educational achievements, marital status could have been a potential and essential factors to analyze their associations with mortality. In addition, manipulation of the dataset was done using R software and Microsoft Excel. Handling missing values led to loss of information since some dataset had to be reshaped in to a wide format in order to access all the variables from a specific column. This transformation automatically recorded and treated some blanks as missing values.

Further studies are needed to examine the associations of factors such as educational achievement, occupational status, pregnancy status. It is important to know the level of association with such factors. Also, data cleaning and handling of missing values when conducting survival analysis should be strictly followed in order to get appropriate results which can be implemented in the society.

5.2 Conclusion

This study provides information about a number of factors and vital signs which are associated with mortality of patients with ebola virus disease. These factors are summarized below:

At the personal level, associations of factors such as sex (female, male) and age groups ($< 40, 41-80, > 80$) where examined. A strong statistically significant association was found between female patients and mortality and also patients within the age group 41-80 had a higher risk of dying than the other younger patients thus highly associated with mortality. At the country level, all the patients from the three West African countries during the 2013-2016 outbreak had a very high risk of death considering their characteristics and vital sings.

In addition, results and interpretations from both multivariable and univariable analysis showed that, some groups of vital signs such as high temperature rate, high respiratory rate, patients with heights more than 1.8m all had a high risk of death thus there are possible factors which are associated with mortality of positive ebola virus patients.

Acknowledgments

I would first like to thank the Lord Almighty for his blessings and for giving me enough strength to be able to finish this essay writing.

I acknowledge my supervisors Dr. Christiana Kartsonaki, Dr. Laura Merson and Dr. Trokon Yeabah for providing me with valuable information, always there to answer to my questions and to guide me through this entire process. Without their support, it would not have been possible to finish this study. I would also like to thank my tutor James Njong Berinyuy for his words of encouragements, his corrections and guides, I am grateful. I would also like to thank the entire staff of AIMS-Cameroon from the Centre President, The Academic Director, through the COO to the logistics and health, the IT manager, all the tutors. Thank you all for being there.

My sincere gratitude goes to my family back home and Abroad. My lovely mother, Mrs. Voilet TENDEKA who would always call to check on my health and progress. My grand aunt Mama Agie for her kind and loving words, my mummy Judith Ndoping who is always there to assist me, my siblings I thank you all.

To my kind hearted friends from Kenya; Alice Wachira and Dorcas Cheboi. Words cannot express my gratitude my friends. Special thanks goes to my professor from the University of Bamenda, Prof. Fouotsa Emmanuel for his constant check ups.

Finally, to my one and only buddy Ajeagah Joy Ataaji who stood by me from day one till the end. Always there to listen and share in my bad and good moments through this process. Thank you everything and May God bless you.

References

[1] Logan Abel, Shiromi M Perera, Derrick Yam, Stephanie Garbern, Stephen B Kennedy, Moses Massaquoi, Foday Sahr, Dayan Woldemichael, Tao Liu, Adam C Levine, et al. Association between oral antimalarial medication administration and mortality among patients with ebola virus disease: a multisite cohort study. *BMC Infectious Diseases*, 22(1):1–9, 2022.

[2] Mahamoud Sama Chérif, Nut Koonrungsesomboon, Diénaba Kassé, Sékou Ditinn Cissé, Saliou Bella Diallo, Fatoumata Chérif, Facély Camara, Eleonor Fundan Avenido, Mandiou Diakité, Mamadou Pathé Diallo, et al. Ebola virus disease in children during the 2014–2015 epidemic in guinea: a nationwide cohort study. *European journal of pediatrics*, 176(6):791–796, 2017.

[3] MS Cherif, N Koonrungsesomboon, MP Diallo, E Le Gall, D Kassé, F Cherif, A Koné, M Diakité, F Camara, and N Magassouba. The predictor of mortality outcome in adult patients with ebola virus disease during the 2014–2015 outbreak in guinea. *European Journal of Clinical Microbiology & Infectious Diseases*, 36(4):689–695, 2017.

[4] Samuel J Crowe, Matthew J Maenner, Solomon Kuah, Bobbie Rae Erickson, Megan Coffee, Barbara Knust, John Klena, Joyce Foday, Darren Hertz, Veerle Hermans, et al. Prognostic indicators for ebola patient survival. *Emerging infectious diseases*, 22(2):217, 2016.

[5] James D Esary, Frank Proschan, and David W Walkup. Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

[6] I Etikan, K Bukirova, and M Yuvali. Choosing statistical tests for survival analysis. *Biom. Biostat. Int. J*, 7:477–481, 2018.

[7] Jean D Gibbons and Jean D Gibbons Fielden. *Nonparametric statistics: An introduction*. Number 90. Sage, 1993.

[8] Mary-Anne Hartley, Alyssa Young, Anh-Minh Tran, Harry Henry Okoni-Williams, Mohamed Suma, Brooke Mancuso, Ahmed Al-Dikhari, and Mohamed Faouzi. Predicting ebola severity: a clinical prioritization score for ebola virus disease. *PLoS neglected tropical diseases*, 11(2):e0005265, 2017.

[9] Stephen P Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

[10] Jia Bainga Kangbai, Christian Heumann, Michael Hoelscher, Foday Sahr, and Guenter Froeschl. Factors associated with length of stay and treatment outcome of ebola patients treated at an ebola treatment center in sierra leone during the peak period of the west african ebola outbreak 2013–2016. *Archives of Public Health*, 79(1):1–14, 2021.

[11] Isra Khawar. Overall and relative survival for cancer patients. Master's thesis, University of Stavanger, Norway, 2019.

[12] Simone Lanini, Gina Portella, Francesco Vairo, Gary P Kobinger, Antonio Pesenti, Martin Langer, Soccoh Kabia, Giorgio Brogiato, Jackson Amone, Concetta Castilletti, et al. Blood kinetics of ebola virus in survivors and nonsurvivors. *The Journal of clinical investigation*, 125(12):4692–4698, 2015.

[13] Jin Li, Hui-Juan Duan, Hao-Yang Chen, Ying-Jie Ji, Xin Zhang, Yi-Hui Rong, Zhe Xu, Li-Jian Sun, Ji-Yuan Zhang, Li-Ming Liu, et al. Age and ebola viral load correlate with mortality and survival time in 288 ebola virus disease patients. *International Journal of Infectious Diseases*, 42:34–39, 2016.

[14] Denis Malvy, Anita K McElroy, Hilde de Clerck, Stephan Günther, and Johan van Griensven. Ebola virus disease. *The Lancet*, 393(10174):936–948, 2019.

[15] Ruvie Lou Maria Custodio Martinez. *Diagnostics for choosing between Log-rank and Wilcoxon tests*. Western Michigan University, 2007.

[16] Tomi Peltola, Aki S Havulinna, Veikko Salomaa, and Aki Vehtari. Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. *BMA@ UAI*, 27:79–88, 2014.

[17] Beyan Y Sana. *Risk Factors Associated with the Contraction of Ebola Virus Disease in Liberia*. PhD thesis, Walden University, 2019.

[18] Lu Tian and Richard Olshen. Survival analysis: Logrank test, 2016.

[19] Xin Zhang, Yihui Rong, Lijian Sun, Liming Liu, Haibin Su, Jian Zhang, Guangju Teng, Ning Du, Haoyang Chen, Yuan Fang, et al. Prognostic analysis of patients with ebola virus disease. *PLoS neglected tropical diseases*, 9, 2015.