MODELLING LENGTH OF HOSPITAL STAY OF IN-PATIENTS TREATED FOR INFECTIOUS DISEASES UNDER ORDINARY CLINICAL CONDITIONS: A COMPETING RISKS MODELS PERSPECTIVE

MSc. (BIOSTATISTICS) THESIS

Louis Masankha Banda

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

JULY, 2012

MODELLING LENGTH OF HOSPITAL STAY OF IN-PATIENTS TREATED FOR INFECTIOUS DISEASES UNDER ORDINARY CLINICAL CONDITIONS: A COMPETING RISKS MODELS PERSPECTIVE

MSc. (BIOSTATISTICS) THESIS

By

Louis Masankha Banda

BSc. (Statistics), University of Malawi

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

July, 2012

DECLARATION

I the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

Louis Masankha Banda

Full Legal Name

_____

Signature:

_____

Date:

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature: _____Date:_____

Mavuto Mukaka, MSc. (Lecturer)

Main Supervisor

Signature: _____Date:_____

Jupiter Simbeye, MSc. (Lecturer)

Programme coordinator

DEDICATION

To the loving memory of my dear father Lawrence Masankha Banda. May his soul continue resting in peace.

ACKNOWLEDGEMENT

My appreciations go to the former coordinator of this Masters programme, Dr. Lawrence Kazembe and the current coordinator Mr. Jupiter Simbeye for a job well done coordinating the classwork which greatly inspired the direction taken in this thesis. Special thanks to my supervisor Mr. Mavuto Mukaka who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way.

I am greatly indebted to the Malawi-Liverpool Wellcome Trust and Queen Elizabeth Central Hospital through Dr. Miguel A. SanJoaQuin for providing the data used in this thesis. My Appreciations also go to the Department of Mathematical Sciences at Chancellor College and all lecturers directly and indirectly involved in the delivery of the coursework modules for this programme. I also thank the members of the staff working behind the scenes supporting this programme. I am forever indebted to you all, God bless.

ABSTRACT

Background: In survival analysis studies the interest is time taken to experience an event of interest. However, the probability of encountering the event of interest is commonly altered in studies where subjects experience an event other than that of interest. The standard survival time analysis methods, such as Kaplan-Meier method and the standard Cox model, fall short of differentiating different causes when competing risks are present. This is overcome by using statistical models that account for competing risks. The aim of the study was to compare and discuss estimates from nonparametric Cumulative Incidence Function, cause-specific hazards and subdistribution hazards in modeling time a patient suffering from infectious diseases spent in hospital until discharged. Death in hospital was identified as a competing risk.

Methods: The nonparametric CIF was applied to the data to estimate the probability that a death or hospital discharge has occurred before a given day. In addition, the cause-specific hazards modeled the effect of HIV status, age and patient's sex in relation to death or being discharged from hospital. The subdistribution hazards which does not assume independence between events was also used to compare results with the cause-specific hazards. Test of assumptions and model diagnostics followed.

Results: Of 829 patients suffering from infectious diseases, 438 (52.4%) were females.452 (54.5%) patients were HIV positive, 116 (14.0%) were HIV negative and 261 (31.5%) had unknown HIV status. The nonparametric CIF, like the rest of models, showed that the HIV positive had a lower probability of being discharged in hospital than the HIV negative. The cause-specific hazard of hospital discharge for males was 0.73 ($p<0.001$). This meant that male patients were 27% less likely to be discharged from hospital compared to females. The subdistribution hazards estimates were close to those by cause-specific hazards. This suggested that the estimation of the hazards of encountering the event discharge was not affected much by the event death.

Conclusions and Recommendation: It is important to follow up cause-specific hazards with subdistribution hazards as it provides a check for the effect competing events on the estimation of probability of occurrence of event of interest. The nonparametric CIF turned out a better estimator of patient's cumulative incidence than the compliment of Kaplan-Meier.

Keywords: *Competing risk, cumulative incidence function (CIF), cause-specific hazards, subdistribution hazards.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

The following are abbreviations and acronyms used in this thesis:

AIDS	Acquired Immuno-Deficiency Syndrome
ART	Antiretroviral Therapy
CI	Cumulative Incidence
CIF	Cumulative Incidence Function
CSH	Cause-specific Hazards
DF	Degrees of Freedom
HIV	Human Immuno-deficiency Virus
HMIS	Health Management Information System
ICD	International Classification of Diseases
ICD 10	International Classification of Diseases version 10
IQR	Interquartile Range
KM	Kaplan-Meier survival estimates
MoH	Ministry of Health
NGO	Non-governmental Organisation
PH	Proportional Hazards
QECH	Queen Elizabeth Central Hospital
SPINE	Surveillance Programme of In-Patients and Epidemiology
WHO	World Health Organisation
$1 - KM$	The complement of Kaplan-Meier survival estimates
95% CI	95 percent confidence interval

CHAPTER 1. INTRODUCTION

1.1. Survival Analysis Method in a Competing Risks Setting

Survival analysis, which is also referred to as time to event analysis, is a class of statistical methods for analyzing data measured from a particular time point until a pre-specified endpoint. In standard survival analysis, an individual who experiencesa pre-specified event of interest within the observation period is said to have an event; otherwise an individual is set to be censored at the end of the study. Participants that encounter events other than that of interest are censored non-informatively. Thus, each study participant makes available two statistics quantities; follow-up time and survival outcome.However, there are other situations where censoring non-informatively the individuals who encounter events other than that of interest alters the estimation of probability of encountering the pre-specified event of interest. Events with such effect on each other are called competing risks.

There are a lot of studies that involve survival time analysis, mostly the standard survival methods which are desperately implemented without even considering the possibility of competing risks among the outcome events. Clinical and Epidemiological investigators sometimes confine themselves only to the statistical methodologies they are familiar with without bothering much to find out first the possibility of engaging other methods which may fully address the objectives of their studies without violating assumptions. Many statistical techniques are based on vital assumptions that must be met before any statistical assessment is completed (Altman *et al*, 1995). Brar (2008) in his published thesis; *Estimation of Cumulative Incidence in the Presence of Competing Risks*, in the literature review found that despite the extensive usage of this method, it is astounding to discover that it is sometimes applied incorrectly or the statistical outputs interpreted inaccurately in the methods section of

some published materials. As a result of this, Brar (2008) concluded that, findings of research studies that misuse survival methods may be deemed questionable.

One of the extensively used methods to estimate survival probabilities is the Kaplan-Meier product limit. The Kaplan-Meier approach provides a nonparametric estimate of the overall survival probability of an event interest (Kaplan and Meier, 1958). Essential to the use of the Kaplan-Meier estimator is the understanding of the concept of censored or incomplete data. Censoring transpires in studies when the exact survival time for subject be followed is not known. The most common type of censoring is right censoring, which indicates the survival time on a subject is incomplete because the subject did not have an event before the end of the patient's follow up in the study. All that is known in the cases is that the survival time exceeds the time of last observation. The underlying assumption of the Kaplan-Meier technique is that censoring of subjects occurs at random; subjects are censored for reasons unrelated to the outcome of the study (Caplan *et al,* 1994). In this case, as Brar (2008) concluded, the probability distribution of survival times for the subjects censored should be comparable to those uncensored. As there is no universally recognized test of random censorship in survival analysis (Brar, 2008), the assessment of this assumption is left to the preference of the analyst, which in many circumstances in a medical research is someone who is not professionally a statistician.

In survival time analysis, the subjects' events can be grouped as either true or cause-specific endpoints.The statisticalimplications for each type of endpoint are not necessarily the same since they are based on different assumptions. Methods of survival analysis are based on the fundamental assumption that all subjects will ultimately fail if the follow up on each subject is complete (Caplan *et al*, 1994). In other words, if a study were tolast a sufficient amount of time it would be possible for investigators to observe an event for each subject. As an illustration, a true survival endpoint include: overall survival where the event of interest is

death from any cause. Methods developed to analyze such data assume the underlying cause for censoring observation is independent to the underlying mechanism for event occurrence (Caplan *et al*, 1994). In principle, this means that subjects who are censored in a study are at equal risk of developing the event of interest compared to those who are still being followed but have not developed the event. This is what is referred to as non-informative censoring.

Study participants at risk of two or more causes of failure are analyzed by methods that allow for competing risks. Kleinbaum and Klein (2005) mentioned that presence of competing risks precludes the occurrence of another event under examination or fundamentally alter the probability of occurrence of this other event. For example the Queen Elizabeth Central Hospital (QECH) adultin-patient data used in this thesis, the interest was to model time until a patient was discharged from the hospital but the competing risk of dying while receive medical treatment precluded the onset of being discharged alive. As a result, subjects who experienced death were not at risk of eventually being discharged alive from hospital. This is a typical example of cause-specific endpointand censoring subjects who develop another event is referred to as informative censoring, which violates a fundamental assumption of the Kaplan-Meier method. Therefore, different methods must be applied for cause-specific endpoints prone to informative censoring. The *Figure 1* illustrates the notion of competing risks where there are up to *k* possible causes of failure.

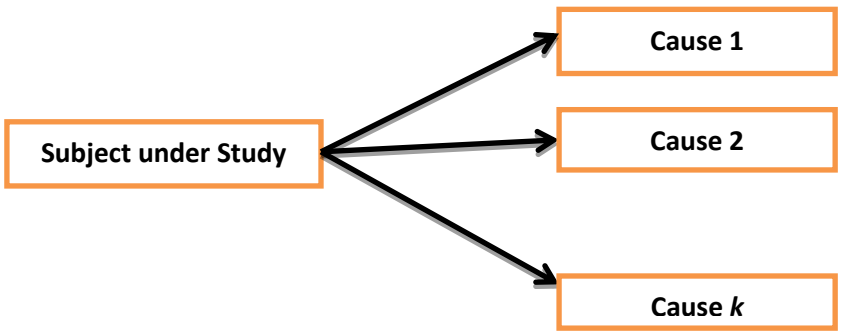

Figure 1: Competing risks situation with k causes of failure.

One of the mathematical definitions of competing risks is related to the joint distribution of time and cause of failure. In the following section the theoretical snippets of this joint distribution are presented.

1.2. Presenting Competing Risk Method as Bivariate Random Variable

Pintilie (2006) in the book *Competing Risks: A Practical Perspective* presented a mathematical way of expressing competing risk method as a bivariate random variable. For each subject the pair (T, D) is observed, where $T \geq 0$ is the time of failure and $D \in \{1, 2, \dots, n\}$ is the failure cause. T is assumed to be continuous and positive random variable while D belongs to exactly one of k different failure types. If an event of type d occurs first, $D = d$, T is then the time at which this event occurred. The joint distribution between T and D is completely specified by either cumulative incidence functions, say $F_d(t)$, or the cause specific hazard function, say $H_d(t)$.

The cumulative incidence functions, CIF, for failure of type d is defined by

$$F_d(t) = \mathrm{P}(T \leq t, D = d)$$

For $t > 0$ and $d \in \{1,\ 2, \dots, n\}$ and corresponds to the sub-distribution function for the probability of failure from cause d in the presence of the competing events.

1.3. Issues with some Survival Analysis Models

When it comes to estimating cumulative incidence the tradition in the past has been calculating one minus the Kaplan-Meier survival probability. Estimating $1-KM$, the failures from competing event are treated as censored at the time this event occurs. This way, the assumption is that the patients failing from a competing risk are no more or less likely to fail from the cause of interest than the patients still at risk beyond this time (Coviello, 2008).

When the aim is to estimate the failure probabilities, this censoring is inappropriate because, after a competing event has occurred, failure from the cause of interest is no longer possible. This is the case since the competing events are assumed to be mutually exclusive.

Kim (2007) also mentioned how the complement of Kaplan-Meier is not an appropriate estimate of cumulative incidence functions. Although $1 - KM$ is conceptually easy to understand and easy to calculate, the estimates are biased if there is more than one type of event and if the events are dependent. This bias arises because the $1 - KM$ method assumes that all events are independent, and thus, censors events other than the event of interest. This type of censoring is what it is referred to as non-informative censoring (Satagopan *et al,* 2004).

The other commonly used competing risk models are the cause-specific Cox models. The cause-specific Cox analysis is applied mostly to explore the pure effect of the covariates. The competing events are censored. As Tai *et al* (2011) put it, cause-specific Cox models are not on their own adequate for modelling competing risk data as such censoring is assumed to be non-informative. The authors also mentioned how this procedure fails to consider that those who have experienced a competing event can never experience the main event of interest. Therefore there is a need to follow up cause-specific Cox model with subdistribution hazards. Only when it has been established that the subdistribution estimates are not different from the proportional cause-specific is the cause-specific very appropriate to fit the data. Lim *et al* (2010) also mentioned that the choice between Cox cause-specific hazards and subdistribution hazards is methodically tailored to the objectives of the study in question.

1.4. Competing Risk Data Assumptions

Underlying this discussion of competing risk data is the existence of two important assumptions. First, it is assumed that the set of *k* competing events are mutually exclusive

and exhaustive. The second assumption is that subjects can experience only one type of event at any particular time point. Models used in competing risk setting come with their own assumption too. The nonparametric cumulative incidence function has the least or no assumptions at all on the data. The semiparametric models do not make any assumption on the shape of the baseline hazards but assumes that it is the same for all events. The parametric models make assumptions on the shape or distribution of the hazard function.

1.5. Objectives of the Study

The intent of this thesis was to identify and evaluate suitable competing risk models of time to discharge from hospital among the adult in-patients suffering from and treated for infectious diseases admitted at the QECH. Death in the hospital was identified as a competing event. The prognostic factors associated with these two outcome events; patient's age, sex, and HIV status were evaluated.

1.5.1. Specific Objectives

1. To compare the estimates probability of failure obtained by the fittingthe complement of Kaplan-Meierand nonparametric cumulative incidence functions; then to also establish whether the probabilities of failurefor males and females, and the HIV positive and the HIV negative were significantly different from each other using the Pepe and Mori test.
2. To interpret and compare the survival hazards obtained from fitting the semiparametric *cause-specific Cox models* and *subdistribution hazard models;* interpret the results and explain possible differences between these the *cause-specific* and *subdistribution hazards* for the QECH spine data.

3. To perform diagnostics with an aim of establishing goodness of fit on the *cause-specific Cox model* and *subdistribution models* fitted to the data and interpret the diagnostics results thereafter.

CHAPTER 2. THE REVIEW OF LITERATURE

2.1. Methodological Issues

Several authors, in the past, expressed concern about the methodological problems coming up in the analysis of cohort studies or clinical trials when competing risks were present (Gooley *et al*, 1999). Investigators would either ignore the competing events by simply doing standard survival analysis (Kim, 2007) or embraced biased estimators of cumulative incidence function. Competing risks occur frequently in cancer research even though their presence may not always be recognized at the time of analysis. As highlighted in the introductory part of their article Coviello and Boggess (2005) defined a competing risk as an 'event whose occurrence precludes or alters the probability of occurrence of main event under examination.' In this setting, the appropriate estimate of the probability of failure is best described by the cumulative incidence. Cumulative incidence of an event is often of interest in medical research and frequently presented in medical articles (Kim, 2007). Previously this had been a huge problem since many statistical software packages could not calculate the cumulative incidence (Gooley *et al*, 1999).

2.2. Handling Cause-specific Endpoints

Cause-specific failure probabilities are used to account the likelihood of a subject failing from a specific event when there is possibility of failing from other events. Methods of estimating cause-specific failure probabilities have been available for quite some time but remain under-utilized in biomedical literature; the reason for this not well known (Pepe *et al*, 1993). A couple of studies have been conducted in Malawi in the public health setting applying competing risk models. One of the most recent studies to apply competing risk model, Weigel *et al* (2012) applied subdistribution to assess the mortality and loss to follow-

up in the first year of anti-retroviral therapy (ART). A great study although cause-specific hazard models were never used. This means that the possibility of independence between the outcomes mortality and loss to follow-up was never assessed. Although the use of competing risk models is gaining ground now, there use is not as high as expected. Perhaps from the technical standpoint, the methods have a predisposition of being mathematically challenging, and deal with the less-than-ideal situation of dealing with more than one event. From the applied side, the reason is most likely lack of awareness among clinical investigators of alternative methods that can be applied.

As one way of promoting the implementation of competing risk models over compromised methods like the complement of Kaplan-Meier estimate, Satagopan *et al* (2004) published a non-technical review, aimed at the applied clinical investigators, recommending and signifying the use of competing risks survival analysis using the cumulative incidence function. The function they explicate was not new in that this is the most common approach to estimate probabilities in the presence of competing risks. A number of authors have examined the estimation of failure probabilities within the competing risks framework. Some of the issues are presented in the following paragraphs.

Gooley *et al* (1999), offered an alternative representation of the cumulative incidence and the complement of the Kaplan-Meier ($1 - KM$) utilizing Efron's concept of reallocating censored observations to the right censored group. They illustrated in their research paper that the $1 - KM$ estimator reallocated competing events to the right censored group in the same way that censored observation were moved to the right, which wrongly assumed that failure from the event of interest was still possible. However, the cumulative incidence estimator removed subjects experiencing the competing event from the risk set and only reallocated the censored observations to the right censored group. Hereafter if a subject failed from the competing

event, the contribution to Gooley's representation of the cumulative incidence is zero. In comparing these two estimators, if no competing risks are available, the cumulative incidence and $1 - KM$ yield exactly similar curves. If there are competing risks, the $1 - KM$ estimate is overblown resulting in biased estimate of failure. Besides, Gooley *et al* (1999) emphasized that the cumulative incidence is founded on the hazard of the event of interest as well as the hazard of the competing risks whereas the $1 - KM$ is just a function of the hazard of failure from the event of interest.

In their paper *Analysis of the Probability and Risk of Cause-specific Failure,* Caplan *et al* (1994) find out that the mechanism of early failure differs from that of late failure in studies involving radiation therapy. For a thorough analysis of local failure the authors advocated the use of the cumulative incidence function. The authors concluded that the cumulative incidence estimator is of particular importance when estimating failure probabilities at a given time but pointed out the estimator failed to convey overall risk for the patient population yet to experience the event of interest. To get over this challenge, they recommended displaying a plot of the cumulative hazard rate, which increased as risk increased but was also difficult to interpret as it lacked a direct probability interpretation. Another approach advocated was tocalculate the cumulative conditional probability.

After estimating the cumulative incidence of an event, it is often of interest to determine whether there is a difference in cumulative incidence rates among different treatment groups. In standard survival analysis, this is done using the log-rank test to compare curves generated via $1 - KM$ method. In the presence of competing risks, however, this is inappropriate, for the same reason given for $1 - KM$. Instead, Kim (2007) cited in his paper that Gray (1988) investigated this issue and proposed a class of tests for comparing cumulative incidence curves of a particular type of failure among different groups in the presence of competing

risks. As cited by many authors in their journal articles, *Pepe* and *Mori* (1993) give a method for comparing the cumulative incidence curves directly. In Stata 10, if you are using a Stata-certified ado file written by Coviello, this method gives out comparison results for both the cumulative incidence curves estimated from event of interest and the curves from competing events (Cleves *et al*, 2010).

2.3. Prognostic Factors in Competing Risk Data Analysis

When the difference in the cumulative incidence curves has been established among different treatment groups, it is also important to determine whether this difference is solely due to treatment or to the confounding factors, such as age. This question is usually fixed by fitting cause-specific Cox model for a particular failure, treating other competing risks as censored (Kim 2007). However, the effect of a covariate on an event from either a cause-specific model may be different from the effect of the covariate of the event in the presence of competing risks (Kim 2007).

Cause-specific hazards and corresponding hazard ratios are estimated using Cox proportional hazards model for each failure event. Cause-specific hazards estimation is most commonly used method of analysis in a competing risk setting (Kleinbaum and Klein, 2005). Cause-specific hazards give insight into the biological mechanism of subject under investigation since they have independent assumption among the competing events.

The comparison of the cause-specific hazards is made as if the other types of events did not exist. This approach is regarded by a good number of investigators as unrealistic (Kim 2007). However, Pintilie (2006) in his book *Competing Risks* stressed that the use of cause-specific hazards is a good way of analyzing the data when one wants to find the biological mechanism underlying the specific outcome.

On the other hand, comparing the cumulative incidence functions is more direct; it takes into account all types of events and does not assume independence between times to the different types of events. However, Pintilie (2006) further argues that the cumulative incidence function for the event of interest can be low just because the risk of a competing risk event is high. On contrary, the cause-specific hazards regression is invariant to the size of the competing risks. Hence, the simple comparison of the cumulative incidence function for the event of interest is not sufficient and needs to be enhanced by the comparison of the cumulative incidence function for the competing risks as well (Coviello and Boggess, 2004).

2.4. The Subdistribution Hazards

Using Cox models alone to model the cause-specific hazards for the event 'hospital discharged' with the covariate say, patient's HIV status, then the resulting cumulative incidence functions for the discharged that assess the HIV status effect will depend on the following five things; (1) the baseline hazard for being discharged; (2) the baseline hazard for dying in the hospital; (3) the effect of HIV status on the hazard for being discharged; (4) effect of HIV status on the hazard for dying in hospital; and finally, (5) time itself. There is no way to summarise how the HIV status affected the incidence of discharged without taking all these factors into account (Cleves *et al*, 2010). Furthermore, with this Cox analysis method you are not even guaranteed that the cumulative incidence for one group will always be greater than that for the other: the curves could cross at one or more points.

Fine and Gray (1999) solved this mystery by proposing a regression modeling applied directly on a cumulative incidence function for a particular use in a competing risks analysis. It is much easier to interpret for cumulative incidence functions. They imposed a proportional hazards assumption on the subdistribution hazards and gave estimators and large samples properties. The subdistribution hazard model is formulated in a similar manner as the cause-

specific Cox model, except that the exponential of the regression coefficients now denote the subdistribution hazard ratios of the respective covariates on the subdistribution hazard of event, say k.

This method takes into account other events and does not make any assumptions about their independence between the event time and censoring distribution. In other words, the censoring mechanism is independent of disease progression. Estimation of the covariates coefficients for the models on cause-specific and subdistribution hazards follows the partial likelihood approach used in the standard Cox model. However, the difference between cause-specific and subdistribution hazards lies in the risk set (Lim *et al,* 2010). For the cause-specific hazards, the risk set decreases at each time point there is an event of another cause. For the subdistribution hazard a person who has an event from another cause remains in the risk set.

2.5. The Admission Data and Competing Risk Models

This thesis applied and compared the performance of $1 - KM$ and nonparametric cumulative incidence; it also compared the cause-specific and subdistribution hazards with an aim of assessing the degree of association between the event of interest and the competing event. Discharged from hospital was the outcome of interest and dying in the hospital was considered as a competing event. The aim was illustrate the implementation of prominent competing risk models often used on epidemiological data and to explore the effect of HIV status, age and sex on the time spent in hospital until discharged. Implementation of competing risk models took care of those who died as having encountered another event. The nonparametric cumulative incidence were applies to estimate overall probability of encountering an event; be it the main event or competing. This model was considered for its very little assumptions it makes on the data. The cause-specific Cox models were applied to

explore the pure effects of individual covariates on the survival time. The subdistribution hazard models were implemented as a semiparametric approach to the cumulative incidence. Unlike cause-specific hazard models, the subdistribution hazard models do not just right censor the competing events when they occur but consider them as another type of events altogether. The subdistribution hazard models also demonstrate the effect of variables by giving out the subdistribution hazards ratios for each variable.

CHAPTER 3. METHODOLOGY

3.1. The Data and Study Population

As clearly highlighted in the preceding chapters, the major interest in this thesis was to model time spent in hospital until a subject was discharged within the observational period of 7 days.

The data used in this thesis was in-patient data collected at the QECH for patients 14 years old or above. Baobab Health Trust, a non-governmental organisation based in Lilongwe, collaborated with the Ministry of Health and Malawi Liverpool Wellcome Trust for deployment of a computerized real time data collection systems to the QECHfor their Surveillance Programme of In-Patients and Epidemiology (SPINE) project. The information system recorded, tracked and managed in-patient care and appointment data. The patient registration system allowed all patients to be recorded with relevant details. Using a unique barcode for each, it was able to identify patients so that their records could be retrieved from system in future visits by simply scanning their assigned barcodes.Having each patient's summary record stored in a computer system meant that whenever a patient was there to seek care from QECH, those treating them would have secure access to summary information to assist with diagnosis and care, and to also know how many times a patient visited the facility. The SPINE data was availed for this thesis in a Microsoft Excel spreadsheet format. It covered patients' diagnosis and admission information from December 2010 to June 2011. As of now, the SPINE data is still being collected on daily basis as part of Health Management Information System (HMIS). This is greatly linked to monitoring and evaluation of the healthcare provided. The interest was to model time the adult in-patients suffering from infectious diseases spent in hospital until discharged. Competing risk models were applied regarding death as a competing event.

The endpoints of time spent in hospital were the health outcomes as listed in the SPINE dataset. There were five outcomes or five ways to end one's hospitalization span; (1) discharged alive (and probably better), (2) dying in hospital, (3) transferred out to a different hospital, (4) referred to another facility, and finally (5) absconding. The main interest was time in days to discharge. Out of the five listed outcomes, only discharge and death occurred frequently hence the other outcomes were ignored as they were very rare events. Only those that died in hospital and those that were discharged produced comparable figures to conduct statistical analysis and were kept within the study population for this thesis. The event of interest being discharged from hospital, death before being discharged was thought as a competing risk event. Only patient's first recorded admission visit was used in the analysis.

The dataset used in this study contained only adult in-patients' information and not any out-patients' information. An adult here was defined as any individual 14 years of age and above. Therefore, the study participants were the admitted adult patients whose information was collected and saved in the SPINE database. Since the QECH is the only public referral hospital in the South-Western Medical Zone, these patients are generally from the districts making up the South-Western Medical Zone. These districts are *Blantyre* itself, *Chiladzulu, Mwanza, Neno, Thyolo, Chikwawa* and *Nsanje*. Patients coming to seek healthcare at QECH for the first time were assigned a spine barcode which comprised a unique number and computer identifiable bars or stripes. This enabled the computer system to identify a patient every time he or she comes to seek healthcare at QECH and allowed each patient's medical history to be collected and stored electronically for reference.

The diseases and disorders recorded through the process of medical diagnosis were too numerous to be statistically considered separately. As such, the diseases and disorders were put into categories as per the international classification of diseases (ICD – 10). The ICD – 10 is a World Health Organisation sanctioned method of putting diseases into groups. After

categorizing the diseases, only patients suffering from the infectious diseases were kept in the dataset for analysis. Concentrating on one disease group ensured that there was statistical homogeneity and that variations due to diseases group type or disorder group type were taken care of. Infectious disease category comprised among others diseases such as all kinds of tuberculosis, sepsis, urinary tract infection, meningitis, and malaria. The reason for settling on infectious diseases was partly that these categories comprised most suffered and reportedly most life threatening diseases in Malawi. It was interesting and important to know length of hospital stay information of patients suffering from infectious diseases and receiving treatment under ordinary clinical conditions at QECH.

The length of time spent in the hospital was measured in days. The entry point into the study was the day a patient was admitted into the hospital and the exit time was the time a patient either died or was discharged from the facility. Since there were competing outcomes in this study, the Cox models were applied with the focus on cause-specific hazards and not standard hazards. For the same reason the cumulative incidence function was opted over the survival function. The estimation of probability occurrence by time, say t, for a particular failure can be handle by fitting $1-KM$, the complement of Kaplan-Meier estimator, or cumulative incidence function. The estimator $1-KM$ was opted out because of bias when dealing with competing events. Nonparametric cumulative incidence function was a better replacement and posed as a rational comparison to the $1-KM$ as both model are purely nonparametric in nature. The cause-specific hazards assume independence among the competing events and are only suitable when biological mechanism of the covariates is of interest. Otherwise they right censor competing events whenever they occur. To overcome this, the subdistribution hazards model by Fine and Grey (1999) were implemented as they recognize a competing event when it has occurred and takes care of competing events when coming up with hazard functions. The other models that can be used in a competing risk setting include the multinomial

logistic. It assumes that the covariate effect is constant across events and assesses whether the baseline hazard is varying across events. The multinomial logit models were not considered for this thesis as they could not help to achieve the objectives set for this study. The following sections present the statistical procedures, relevant mathematical characteristics of the models in this thesis.

3.2. Standard Single Event Time Model

In follow-up studies the exact survival time is only known for those study participants or units who show the event of interest during the follow-up period. For the others all one can say is that they did not show the event of interest during the follow-up period. These study participants or units are called censored observations. Individuals can be right censored, left censored or interval censored. Subjects are right censored if it is known that the event of interest happened sometime before the recorded follow up time (Kleinbaum and Klein, 2005). An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set.

Standard survival data measure the time span from some time of origin until the occurrence of one type of event. In such a case, the Kaplan-Meier product limit estimator is frequently used in describing time to event experience of the subjects under study. The standard survival data can also be presented as a bivariate random variable, say (T, D), where $T \geq 0$ is time to event of interest and $D \in \{0,\ 1\}$ is the failure cause. D here is the censoring variable, $D = 1$ if the event of interest was observed, and $D = 0$ if the observation as censored. When $D = 1$, then T is the time at which the event occurred and when $D = 0$ is the time at which the observation was censored.

In general as Pintilie (2006) put it, given T as a random variable representing survival time that has a density function, $f(t)$, and distribution function, $F(t)$. The survival function at

time t, $S(t)$, is defined to be the probability that the survival time is greater than t, where $S(t) = P(T > t) = 1 - F(t)$. The survival function, therefore, represents the probability that an individual survives from the origin to sometime beyond t. The hazards function or hazard rate, $h(t)$, is the probability that an individual encounters an event of interest at time t, conditional on having survived to that time, which is defined as:

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right\}$$

The hazard function, therefore, represents the instantaneous death rate for an individual surviving up to time t and provides full characterization of the distribution of T.

The main concern with this approach is how to study the impact of covariates of the distribution of T. To do this, we assume the variation in the distribution of event and censoring can be characterized by a vector of observed explanatory variables, say Z, which can be either time-invariant or time-dependent covariates. Under Cox proportional hazards model, the hazard function for the event time T associated with the covariates Z is defined as follows:

$$h(t) = h_0(t) e^{\beta' \mathbf{z}}$$

Where the function $h_0(t)$ is an unspecified baseline hazard function and gives the shape of the hazard function. If all explanatory variables are zero, the hazard function will be the baseline hazard $h_0(t)$. If two individuals have identical values of the measured covariates, they will have identical hazard functions.

3.3. Cause Specific Hazards Models

Again as stated in the previous chapters, competing risks in survival analysis refer to a situation where subjects under investigation are exposed to more than one possible type of

events. Thus, each subject is associated with a pair (T, D) where T is the time to event and $D \in \{1, 2, \dots, k\}$is the type of the event for that subject (Pintilie, 2006). In this case there are k possible causes of failure. The cause-specific hazard function in the competing risks model is the hazard of failing from a given cause k in the presence of the competing events as shown mathematically below (Kleinbaum and Klein, 2005):

$$h_k = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t} \right\}$$

With $D \in \{1, 2, \dots, k\}$. With covariates incorporated in it, the regression model on cause-specific hazards is given as:

$$h_k(t|z) = h_{0k}(t) e^{\beta' z}$$

The total hazard, $h(t: z)$, equals the values of its corresponding hazards function summed up to time t. It is then

$$h(t|z) = \sum_{k=1}^{k} h_k(t)$$

This equation implies that the all-cause hazard rate is the sum of K hazards.

The cause-specific hazard can be modeled using the Cox model, which is broadly used in medical research. The cause-specific hazard model may be more clinically understandable when assessing the prognostic effect of the covariates on a specific cause because it can be observed whether the covariate is reducing or increasing the instantaneous probability of the event of interest irrespective of other covariate effect.

3.4. Nonparametric Cumulative Incidence Function

With competing risk data, the cumulative incidence curve derived from cause-specific hazard functions provides important event information for a specific cause. Marubini and Valsecchi (1989) derived the cumulative incidence estimator for the failure k as

$$\hat{I}_k = \sum_{j|t_i \leq t} \hat{S}\left(t_{j-1}\right) \frac{d_{kj}}{n_j}$$

Where $\hat{S}(t_{j-1})$ is the Kaplan-Meier estimate of the overall survival function, that is, considering failures of any kind, and the second factor is an estimate of the hazard of failure k. This definition implies that the cumulative incidence is a function of the hazards of all the competing events and not solely of the hazard of the event to which it refers. This equation further shows that the sum of all cumulative incidences equals $1 - \hat{S}(t)$, the complement of the overall Kaplan-Meier estimate of survival considering failures of any kind.

The variance estimator for the distribution of this formula is as follows (Caviello and Boggess, 2004):

$$Var\{\hat{I}_k(t_j)\} = \sum_{\alpha=1}^{j} \left[\{\hat{I}_k(t_j) - \hat{I}_k(t_\alpha)\}^2 \frac{d_\alpha}{n_\alpha(n_\alpha - d_\alpha)}\right] + \sum_{\alpha=1}^{j} \{\hat{S}(t_{\alpha-1})\}^2 \left(\frac{n_\alpha - d_{ka}}{n_a}\right)\left(\frac{d_{ka}}{n_\alpha^2}\right)$$

$$-2\sum_{\alpha=1}^{j} \{\hat{I}_k(t_j) - \hat{I}_k(t_\alpha)\}\{\hat{S}(t_{\alpha-1})\}\left(\frac{d_{k\alpha}}{n_\alpha^2}\right)$$

Where $d_j = \sum_{k=1}^{C} d_{kj}$ and C is the number of causes of failure. It was report by Caviello *et al* that a general formula was derived by Dinse and Larson (1986) using the delta method.

3.5. A Comparison of Cause-Specific Hazards Regression and Cumulative Incidence Function

The probability that the event occurs before time t can be derived from the hazard through an equation. So, the hazard completely describes this probability distribution. The higher the hazard, the higher the probability that the event occurs before t and vice versa.

In competing risk situation, the probability that the main event occurs before time t (cumulative incidence) depends on both the hazard of the main event and the hazard of the competing event. Thus, there is no obvious relationship between the hazard and the cumulative incidence of the main event, the latter depending on the hazard of the competing event too.

In competing risk situation the cause-specific hazard and the cumulative incidence do not convey the same pieces of information. The former tells about the biological mechanism underlying the specific outcome. The latter informs us about the probability and, therefore, the actual number of patients failing from a specific cause, taking into account that this type of event could not have been observed, hindered or precluded because of another type of event.

3.6. Subdistribution Hazards Regression

Fine and Gray (1999) developed an alternative semiparametric model that considers all important factors in a competing risk setting. These factors are the baseline hazard effect for the outcome events, the covariate effect for the outcome events and the effect of time itself. They define the subdistribution function for failure cause i as

$$\bar{h}_i(t) = \frac{p\{t \leq T < t + \Delta t,\ failure\ from\ cause\ i | T > t\ or\ (T \leq t\ and\ not\ cause\ i)\}}{\Delta t}$$

This means that the subdistribution hazards for cause i is the instantaneous probability of failure from cause i at time t given either no failure before t or failure from another cause before t. The subdistribution function appeal arises from the fact that the cumulative incidence function for a particular cause i is a function of the subdistribution hazard only for cause i. Mathematically this can be presented as;

$$CIF_i(t) = 1 - exp\left\{-\int_0^t \overline{h_i}(u)du\right\}$$

Where the integral on the right is the cumulative subdistribution function, $\overline{H_i(t)}$. In other words, if you define a regression model for $\overline{h_i}(t)$, you can use it directly to directly interpret covariate effects on $CIF_i(t)$ because there is a direct correspondence between the two.

The Fine and Gray model is a direct analog to Cox regression with the subdistribution hazards taking the place of traditional hazards functions. Their model for subdistribution hazards for cause I is

$$\overline{h_i}(t|x) = \overline{h_{i,0}}(t)exp(x\beta)$$

For covariate vector $\boldsymbol{X}$ and baseline subdistribution hazard function $\overline{h_{i,0}}(t)$. As in Cox regression, this model is semiparametric in that we assume no functional form for the baseline subdistribution hazard. The effects of covariates are assumed to be proportional too.

3.7. Time Varying Covariates

Kleinbaum and Klein (2005) defined time varying covariate as any covariate whose value for a given subject may differ over t. In contrast, a time-independent variable is a variable whose value for a given subject remains the same over t. The general form of the Cox proportional hazards model, as presented earlier, is as follows:

$$h(t|X) = h_0(t)e^{X'\beta}$$

This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variable vector $\boldsymbol{X}$. The Cox model formula says that the hazard at time t is the product of two quantities; the baseline hazard $h_0(t)$ and the exponential expression $e^{X'\beta}$. An important feature of this formula, which concerns proportional hazards assumption, is that the baseline hazard is a function of t but does not involvethe X's, whereas the exponential expression involves the X's but does not involve t. In this case the X's are called time-independent covariates.

There is a possibility, nevertheless, to consider $\boldsymbol{X}$'s that do involve t. If time-dependent variables are considered, the Cox model form may still be used, but they no longer satisfy the proportional hazard assumption. These models are commonly referred to as the extended Cox models.

When time-dependent variables are used to assess the proportional hazard assumption for a time-independent variable, the Cox model is extended to contain interaction terms involving the time-independent being assessed and some function of time.

3.8. Checking the Model Assumptions and Diagnostics

Basically, there are three types of models considered in this thesis. These are nonparametric cumulative incidence function, cause-specific hazards Cox, and subdistribution hazards model. Each one of these was fit taking into account the assumptions that the model make on the data.

3.8.1. The Proportional Hazards Assumption

The Cox proportional hazard model assumes that the hazard ratio comparing any two specifications of the predictors is constant over time. This also means that the hazard for one individual is proportional to the hazard for any other individual, where proportionality constant is independent of time (Cleves *et al,* 2010).

As Cleves *et al* (2010) put it; the Cox model formula says that the hazard at time t is the product of two quantities. The first of these is the baseline hazard function, which is only a function of t and does not involve the explanatory vector $\boldsymbol{X}$. The second quantity is the exponential expression e to the linear sum of $\beta_i x_i$ where the sum is over the k explanatory X variables. The exponential expression does not involve t. The proportional assumption is not met if the graphs of the hazards cross for two or more categories of a predictor of interest. However, as put by Kleinbaum and Klein (2006), even if the hazard functions do not cross, it is possible that the proportional hazards assumption is not met.

3.8.2. Goodness-of-Fit

Model diagnostics are applied to identify unexpected characteristics of the data that may seriously influence conclusions or require special attention. The detection of influential observations, that is observation whose deletion, either singly or multiply, result in substantial changes in parameter estimates, fitted values or tests of hypothesis.

Diagnostic methods are generally based on residuals. Standardized residuals will be produced and assessed, where each residual is standardized by its estimated standard error. Another way of doing this is by correcting for the leverage of the point in the space of the explanatory variable.

3.9. Statistical Software Package

The dataset was obtained as a single document in Microsoft Office Excel with eight separate working sheets. The data on these working sheets were cleaned and exported to a different statistical package called Stata® version 10.0. In general, Stata® is powerful, *interactive* and user-friendly software with high level applicability in inferential statistics. It has to be mentioned here that under ordinary circumstances Stata® 10 cannot handle nonparametric competing risk models. For example, the nonparametric estimation and testing of cumulative incidence functions requires that one download and install some Stata certified user-written software, which provide functionality not included in the official Stata® 10.

To conduct the thorough analysis on nonparametric cumulative incidence functions, there were basically two extra programs needed from Statistical Software Components archives hosted at Boston College in the United States. To estimate nonparametric cumulative incidence functions, there was a need to first install the command `stcompet` by Coviello and Boggess (2004). To test equality of cumulative incidence functions among groups, the command `stpepemori` written by Coviello (2008) was installed. The subdistribution hazards were performed using Stata 11 command `stcrreg` which is also not available in Stata 10.

3.10. The Estimates, Statistical Tests and the Level of Significance

The summary characteristics of patients such as age and average days spent in hospital were presented as median and, naturally, the measure of dispersion was the interquartile range. Rank-based measure of central tendency and its subsequent measure of dispersion are ideal in survival data since survival data are typically right skewed. The Hazard ratios, their corresponding coefficients and 95% confidence intervals were presented for cause-specific

hazards and subdistribution hazards models. Also included were the calculated p-values for all statistics. All statistical tests were made at 5% level of significance.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter presents and discusses the results. Section 4.1 presents exploratory data analysis, the fitted models are presented in section 4.2, the assessment of model assumption and goodness-of-fit is outlined in section 4.3. Finally, section 4.4 presents the discussion of results.

4.1. Exploratory Data Analysis

The SPINE dataset constituted 7262 patients who were admitted at QECH between December 2009 and June 2011 for different diseases. Of the 7262 patients, only 829 met the ICD – 10 criteria as suffering from infectious diseases and were therefore included for analysis. *Table 1* below shows the baseline characteristics of the patients.

Table 1: A summary of characteristics of patients

Characteristics N=829		HIV Status			Sex	
		HIV+	**HIV-**	**Unknown**	**Male**	**Female**
Patient's Age (Years)	Median	34.1	30.6	32.1	34.1	31.6
	IQR	12	21.3	17.6	14.8	13.5
Time in hospital (Days)	Median	7	5	3	5	4
	IQR	3	4	4	4	5
Discharged Alive n=486	Frequency (percent)	222 (45.7)	70 (14.4)	194 (39.9)	206 (42.4)	280 (57.6)
Died in Hospital n=82	Frequency (percent)	41 (50.0)	3 (3.7)	38 (46.3)	45 (54.9)	37 (45.1)
Censored at the end n=261	Frequency (percent)	189 (72.4)	43 (16.5)	29 (11.1)	140 (53.6)	121 (46.4)

The summary of baseline characteristics as presented in *table 1*, out of 829 patients, 438 (52.8%) were females. 452 (54.5%) patients were HIV positive, 116 (14.0%) were HIV negative and 261 (31.5%) had unknown HIV status. The overall median age for 829 patients was 35.3 years with an IQR of 21.0. The median ages for male patients and female patients were close with very similar dispersion, males had a median age of 34.1 years (IQR: 14.8) and females had median age 31.6 years (IQR: 13.5). For the HIV positive patients, the median age was 34.1 years (IQR: 12.0) and the HIV negative patients' the median was 30.6 years (IQR: 21.3). From the interquartile ranges, it was clear that the HIV negative patients' ages were much more dispersed than the HIV negative patients' ages. The HIV positive patients had a relatively small interquartile range which signified that their population was concentrated around the maiden age 34.1 years.On the health outcomes of the patients; 702 (84.3%) were discharged alive from the hospital, 127 (15.7%) were reported to have died in hospital while receiving medical treatment.

Infectious diseases category constituted a wide range of diseases most commonly include different kinds of tuberculosis, urinary tract infection, sepsis, and malaria. Some of the admitted patients were also diagnosed with other infectious diseases such as hepatitis, meningitis, and measles. But there occurrence rate was relatively low henceforth bundled up in a sub-group called 'other'. *Table 2* presents the diagnosis results of infectious diseases as defined by the WHO sanctioned ICD – 10:

Table 2: A summary of clinical diagnoses results performed on the in-patients

Disease $N = 829$	**Frequency (Percentage)**
Tuberculosis	188 (22.7)
Urinary Tract Infection	43 (5.2)
Malaria	223 (26.9)
Sepsis	307 (37.0)
Other *(hepatitis, meningitis, measles, etc.)*	68 (8.2)

The outlying ages are also evident by segregating by HIV status. *Figure 2* shows Box-plot diagrams showing the dispersion of age among the observations, categorized by sex and HIV status:

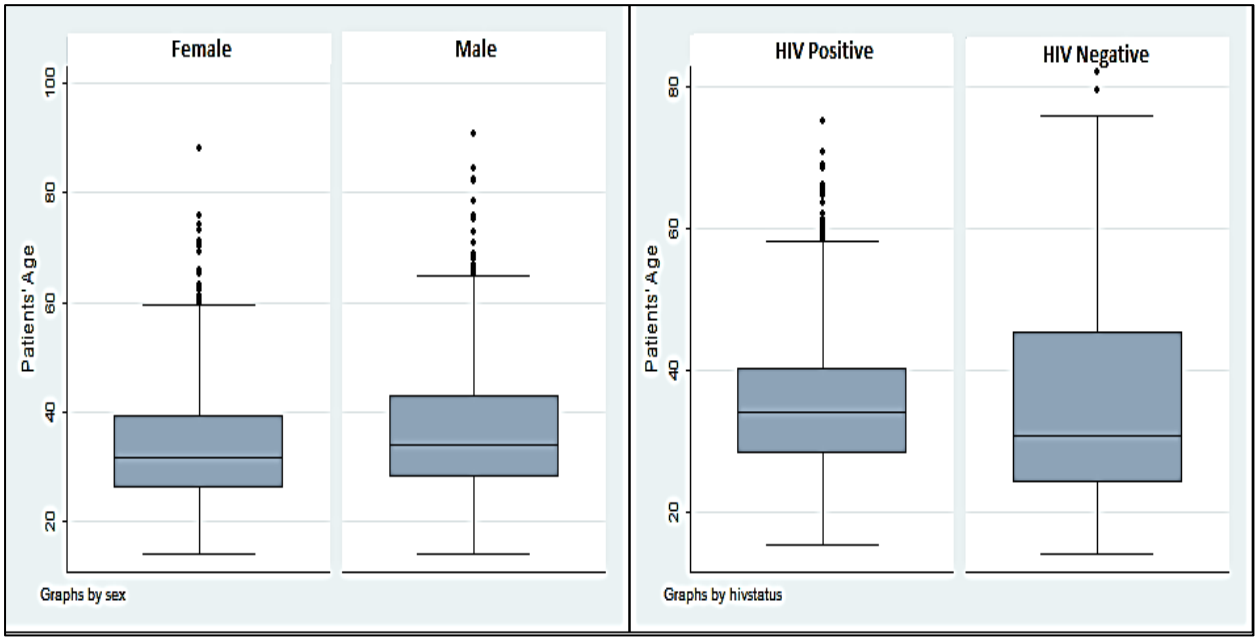

Figure 2: The Box-Plot by sex and by HIV status.

The presence of age outliers, as shown by the Box plots in *figure 2*, was confirmed. Both female patients and male patients had outlying aged patients. The HIV status Box-plots shows that the distribution of the HIV positive patients is concentrated between ages 30 and

40, which contrasts sharply with the Box-plot of the HIV negative patients. The concentration of age distribution of the HIV negative is spread out across twenties to fifties. For the admitted HIV positive patients, this meant that prevalence was relatively high among those between late twenties and late thirties.

Table 3 shows the patients enrolled as suffering from infectious disease from December 2009 to June 2011. The lowest observed enrollment rate was 10 and that was in December 2009. The highest observed enrollment rate was 71 in January 2011. For the other months, the number of patients registered varied between these two highest and lowest figures. In the last month June 2011 a total of 20 patients were registered. The entire data collection span consisted of 19 months. The following table shows how patients' admissions were distributed across months and years from December 2009 to June 2011, the data collection span for this dataset:

Table 3: Patients admitted at QECH for generally suffering from regular infectious diseases

Month Admitted $N = 829$	**Year Admitted**		
	2009	**2010**	**2011**
January	-	29	71
February	-	59	50
March	-	66	36
April	-	51	25
May	-	45	31
June	-	17	20
July	-	40	-
August	-	42	-
September	-	50	-
October	-	67	-
November	-	65	-
December	10	55	-

Table 4 is a life table showing the survival pattern of the patients within the observation time of 7 days. From *this table*, 486 patients were discharged within the observation period of 7 days in hospital. Of the 829 patients, 361 (43.6%) of them remained in the study after 5 days, the rest either died or discharged. On the seventh day, only 316 patients and were henceforth censored.

Table 4: Shows the estimated survival probabilities of the days of the patients admitted at QECH.

Observation Time (Day)	**Beginning Total**	**Discharges in Time Interval**	**Estimated Survival Probability**	**Standard Error**
1	829	126	0.85	0.012
2	703	95	0.73	0.015
3	608	89	0.63	0.017
4	519	88	0.52	0.017
5	431	70	0.44	0.017
6	361	45	0.32	0.017
7	316	316	0	-

In this analysis, no participant was reported lost to follow up. The estimated survival probabilities given in the table above were calculated under standard survival assumption where subject who died in the hospital were treated as censored observations.These estimates would be biased if the event of interest and the competing event were dependent.

4.2. The Models Fitted

This section presents the cumulative incidence function, cause-specific hazards, subdistribution hazards models that were applied to the SPINE data and their statistical inference implications.

4.2.1. The Comparison between Nonparametric Cumulative Incidence and $1 - KM$

The results obtained after fitting the nonparametric cumulative incidence function were compared to those of the complement of Kaplan-Meier $1 - KM$. As shown in the *figure 4-2*, the estimates of $1 - KM$ were deviating far and far away from those obtained from the nonparametric cumulative incidence with time. The independence of competing events assumption made for $1 - KM$ was clearly not valid as the curves were slowly deviating apart with time. In this scenario, it was important to treat death as another event, a competing event. Since the nonparametric cumulative incidence considers both the main and competing event when plotting cumulative survivorship curve, it was clear from *figure 3* that the two events were dependent on each other to some extent. *Figure 3* shows the $1 - KM$ and nonparametric cumulative incidence curves.

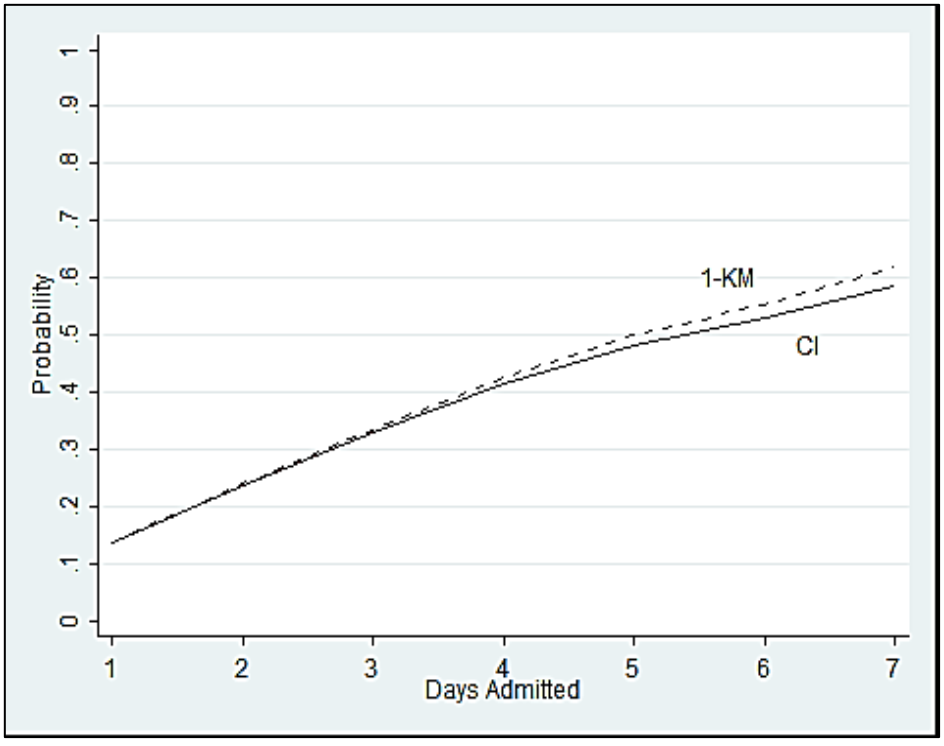

Figure 3: Comparison of $1 - KM$ and Cumulative Incidence (CI) curves.

From *figure 3*, at time 1 both the cumulative incidence and the $1-KM$ model gave similar survival probability. As the admission days progressed the $1-KM$ curve gave higher estimates as it censored those who encountered death in the course of admission. Therefore the cumulative incidence function was a safer model to opt for. However, it was noted in *figure 3* that the estimates from $1-KM$ and cumulative incidence function were not hugely different due to the fact that by far more patients were being discharged than dying in hospital.

4.2.2. The Comparison of Cumulative Incidence Functions between Males and Females, and between the HIV Positive and HIV Negative

The nonparametric estimation of the cumulative incidence functions for groups were plotted, comparing the survivorship of males and females, and the survivorship of HIV positive and the HIV negative patients across all ages. *Figure 4-3* shows cumulative incidence curves for the males and females.This was the calculated probability of being discharged from hospital given that others died along the way.

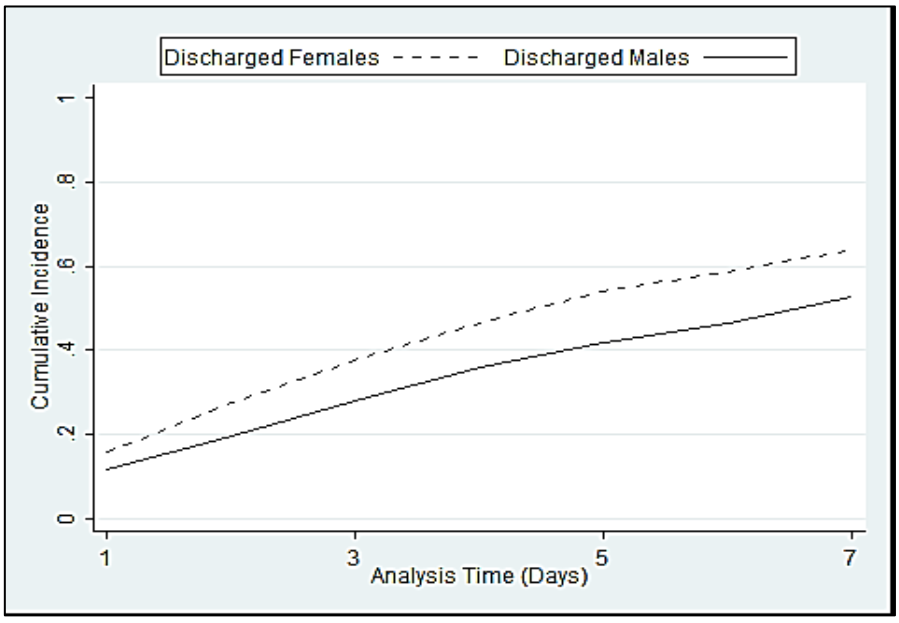

Figure 4: Cumulative Incidence by sex for the outcome 'discharged from hospital'.

The cumulative incidence curves from *figure 4*; the curve for the females was consistently higher than that of males. This meant that, for whatever reasons, female patients had a consistent higher likelihood of being discharged from hospital than their male counterparts. On day 1, both cumulative incidences were below 0.20. The females' cumulative incidence was approximately 0.15 and that of males was approximately 0.10. On the seventh day, the estimated cumulative incidence for females was 0.62 and that of males was approximately 0.52.

The overall cumulative incidences for both events showed that patients were more likely to be discharged than die in the hospital. Of course the cumulative incidence curves were disaggregated by HIV status but the overall incidences can be discerned from that.

Figure 5 shows the HIV status cumulative incidences graphed by HIV status. The first graph shows the cumulative incidence when the event is a patients being discharged from the hospital, and the second graph shows cumulative incidence when death was the event that occurred. The cumulative incidencefor the HIV negative patients was higher than of HIV negative patients, this meant that the HIV negative were more likely to be discharged from hospital as compared to the HIV positive patients.On death as an outcome event, the HIV positive were by far more likely to die in the hospital than the HIV negative patients. This obviously meant that patient's HIV status had an effect on the health outcome events of a patient admitted at the QECH.

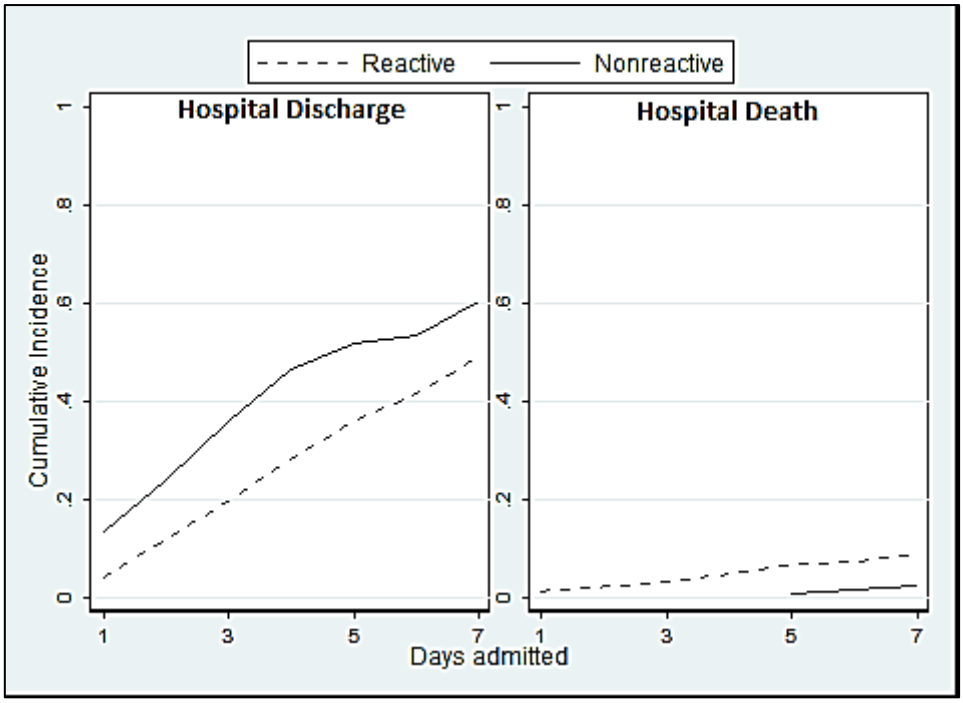

Figure 5: Cumulative incidence by HIV status. The first one is for the failure event of being discharged; the second is for the event 'death in hospital'.

Table 5 presents the figures obtained after running Pepe and Mori cumulative incidence comparison test for both the competing and the event of interest.

Table 5: The measurements after applying Pepe and Mori cumulative incidence comparison test

Extrapolative Factor	**Outcome Event**		**Chi-Square (1)**	**P-Value**
Patients' Sex	***Main Event***	Discharged	14.13	$p < 0.001$
	Competing	Died	24.20	$p < 0.001$
HIV Status	***Main Event***	Discharged	13.81	$p < 0.001$
	Competing	Died	0.61	0.435

The first three p-values obtained for both events lead to the rejection of the null hypothesis that the cumulative incidences were similar. The p-values were all less than 0.001. For those discharged from hospital, it was therefore concluded that there was enough evidence that the male and female cumulative incidences were statistically different from each other. It was also concluded that the cumulative incidence curve for the HIV positive was significantly differently from that of HIV negative. There was association between patient HIV status and the likelihood of being discharged from hospital. From the cumulative incidence curves plotted, it is quite distinct that the HIV negative patients had a consistently higher likelihood of being discharged from QECH as compared to their HIV positive counterparts. Putting this in terms of length of hospital stay, the HIV positive patients seemed to have been spending a little more time in the hospital before being discharged as compared to the HIV negatives.

The cumulative incidence by HIV status for the competing event death in hospital came out insignificant with a p-value of 0.435. There was no enough evidence to reject the null hypothesis and it was concluded that the cumulative incidences were not statistically different from each other. Interpreting this further in terms of length of hospital stay, there was gross lack of evidence about the difference in probabilityof spending time in hospital before a patient was died between the HIV positive and the HIV negative.

4.2.3. The Results for the Unadjusted Cause-Specific Hazards for the Discharged Patients

Three modelswere fitted each containing one of the three covariates; these being HIV Status, age and patient's sex.*Table 6* shows the coefficient and hazard ratio estimates gotten after fitting three cause-specific hazard modelsfor each covariate HIV status, age and patient's sex. The reference category for HIV status was the HIV positive. The reference category for sex was the females. As for patients' age, the comparison was based on per unit increase in age.

The failure event for these models was the event of interest, discharged from hospital. Death in hospital was treated as censored observation. All models were unadjusted.

Table 6: The coefficient estimates after fitting three unadjusted Cause-Specific Hazard (CSH) models with 'discharged' as the failure event.

CSH Model	**Hazard Ratio**	**95% CI**	**P-value**
HIV Negative (Reference: HIV Positive)	1.40	1.07 , 1.83	0.014
Age	0.99	0.98, 0.997	0.007
Male patients (Reference: Females)	0.74	0.62, 0.88	0.001

From this output, the estimate of the hazard ratio for the HIV negative patients as shown in the cause-specific model was 1.40 *(95% CI: 1.07, 1.83).* The p-value for the Wald test was for this was 0.014 which is less than 0.05. Thus HIV negative patients had a 40% higher hazard of encountering hospital discharge than their HIV positive counterparts. Put differently according to the figures presented in *table 6,* the HIV positivepatients had a lower hazard of encountering the event of interest. The cause-specific hazard modelfor patient's age yielded a hazard ratio estimate of 0.99 *(95% CI: 0.98, 0.997)* and a p-value of 0.007. Since the hazard ratio was less than 1 and age progressed by years; this meant that any patient one year older had a 1% less hazard of being discharged from hospital as compared to a patient a year less in age. Patient's sex came out significant too with the hazard ratio of 0.74*(95% CI: 0.62, 0.88)*and p-value of 0.001. Males had a 27% less chance of being discharged from hospital compared to females. *Figure 6* presents the graphical regression results for cause-specific hazard Cox model with HIV status as treatment variable among the patient.

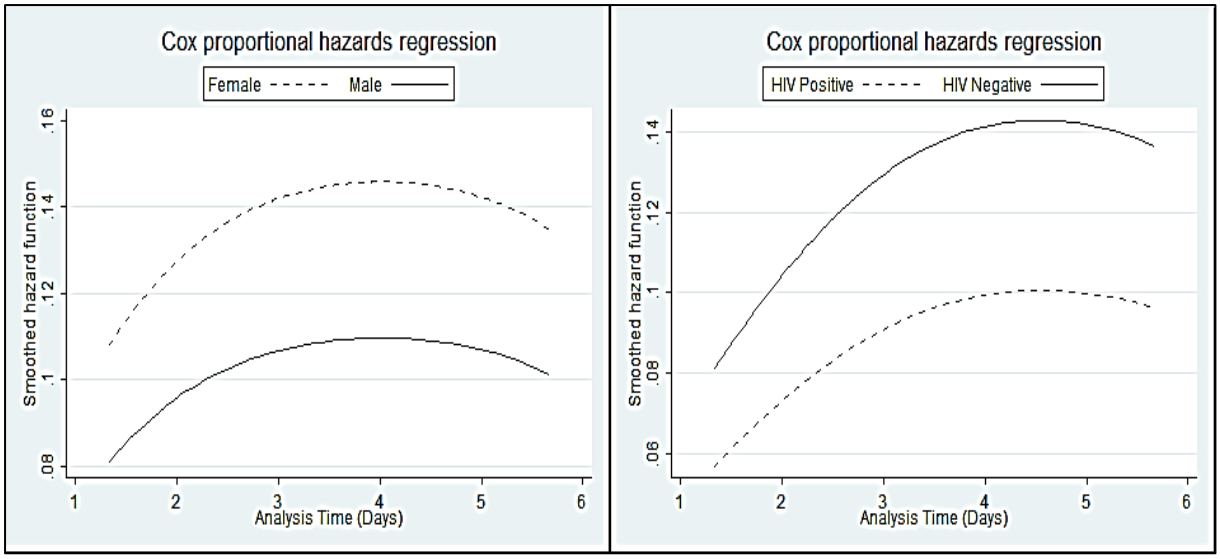

Figure 6: Smoothed cause-specific hazards curves when 'discharged' is the failure event.

Figure 6 confirms the results from *table 5*; females had a higher hazard of encountering the event of interest than their male counterparts, same with the HIV negative patients who had a higher hazard of being discharged. As time a patient spent in hospital was increasing so were the hazard curves for both the patient's HIV status and patient'ssex, until towards the very end where the hazard curves seem to lower.

4.2.4. The Results of Unadjusted Cause-Specific Model for the Competing Event Death

The same three unadjusted cause-specific hazards modelswere fitted with failure event type as death in hospital. Here, now discharged patients were treated as censored observations. By censoring the discharged patients, it was assumed that the event discharge was independent from death in hospital. All models we fitted unadjusted, the results after fitting these models are presented in *table 7*:

Table 7: Estimates of unadjusted cause-specific hazard (CSH) Cox models with death as failure event

CSH Model	**Hazard Ratio**	**95% CI**	**P-value**
HIV Negative (Reference: HIV Positive)	0.33	0.10, 1.05	0.061
Age	1.02	1.003, 1.03	0.018
Male Patients (Reference: Females)	1.20	0.78, 1.85	0.418

The cause-specific hazard model for age is the only one that came out significant when failure event was 'death in the hospital'.The p-value was 0.018 and hazard ratio was 1.02 *(95% CI: 1.003, 1.03)*. This meant that patients a year older yielded a 2% higher hazard of encountering death in hospital than a year younger patients. Simply put; old patients had a higher likelihood of dying in hospital than the young patients. Both the cause-specific hazard models for the factors sex and HIV status were insignificant. This implied that, under independence of outcome events assumption, HIV status and patient's age did not have a significant impact on the hazard of encountering death in hospital.

4.2.5. Fitted Adjusted Cause-Specific Hazard Models for the Competing Event Death

After fitting ordinary cause-specific hazard model, it was thought to still explore further the effect of these covariates after setting other covariates constant or adjusting for the other covariates. The *table 8* show the estimates obtained after fitting adjusted cause-specific hazards models. From *table 8* in the HIV status adjusted model, patient's HIV status was now significant with a p-value of 0.036. The HIV negative patients had 72% less cause-specific hazard of encountering death in hospital than the HIV positive. Being an adjusted model, this was in context that age was constant over time and sex was similar. Patient's age was significant again with a p-value of 0.042. But patient's sex was not significant. The hazard ratio for HIV negative over HIV positive patients was 0.28.

Table 8: Adjusted cause-specific hazard (CSH) Cox models with death as failure event

Adjusted Models	**Hazard Ratio**	**95% CI**	**P-value**
HIV Negative (Reference: HIV Positive)	0.28	0.09, 0.92	0.036
Age	1.02	1.001, 1.05	0.042
Male Patients (Reference: Females)	1.28	0.70, 2.36	0.426

The same interpretation can be extended to adjusted cause-specific hazard age model. Age as a covariate was significant with a p-value of 0.042, so was the factor HIV status with a p-value 0.036. Sex was insignificant again. The hazard ratio for age was 1.02 *(95% CI: 1.001, 1.05)*. This meant that every time age was a year higher, the hazard of encountering death as an outcome event increased by about 2%. This implied that with a unit increase in age the hazard of encountering death in hospital increased by 2% among the patients. Since the hazard is usually positive correlated with the probability of encountering the event, the older patients were more likely to die in hospital than the youngerpatients.

The adjusted cause-specific hazards model for patient's sex had patient's sex itself not significant as a covariate. Because of that, no interpretation was made on it.

The cause-specific hazards were followed up by a graphical visualization of the hazard curves for the HIV positive patients and HIV negative patients when 'death in hospital' was failure event. *Figure 7* shows the smoothed cause-specific hazard curves for patients' HIV status.From *figure 7*, the cause-specific hazard curves share similar shapes but the hazard curve for the HIV positive is way above the hazard of the HIV negative. The hazards are at their highest points between the fifth and sixth days.

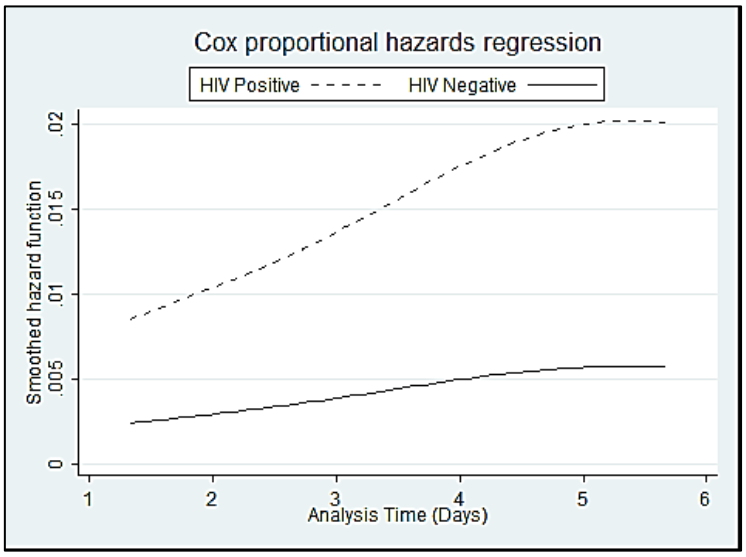

Figure 7: Smoothed cause-specific hazard function for HIV status when the failure event was 'death'.

4.2.6. The Results for the Subdistribution Hazard Models

The last models to be fitted were the subdistribution hazard models. They provide a good check for the independence of events assumption made when implementing cause-specific hazards models. *Table 9* shows the unadjusted estimates yielded after fitting three

subdistribution hazard models for each covariate for the failure event discharged. The results for the cause-specific model are also presented for comparison purposes:

Table 9: Estimates of subdistribution and cause-specific models for discharged event

Model Type	**Models**	**Hazard Ratio**	**95% CI**	**P-value**
Subdistribution Hazards	*HIV Negative* (Reference: HIV Positive)	1.47	1.13, 1.91	0.004
	Age	0.98	0.97, 0.996	0.002
	Male Patients (Reference: Females)	0.74	0.62, 0.87	$P < 0.001$
Cause-Specific Hazards	*HIV Negative* (Reference: HIV Positive)	1.40	1.07, 1.83	0.014
	Age	0.99	0.98, 0.997	0.007
	Male Patients (Reference: Females)	0.74	0.62, 0.88	0.001

The results show that the effect sizes from the cause-specific and subdistribution hazards models were pretty much close for ’discharged’ event. This can be confirmed by comparing the corresponding hazard ratios for both subdistribution and cause-specific hazards models. This meant that the effect on the hazards from competing risk death was quite minimal. This is loosely consistent with the assumptions made when implementing cause-specific hazards that there is no trait effect on the hazard from competing events. In a scenario where cause-specific hazards and subdistribution hazards are similar or close, the cause-specific hazards model would be just enough.

The subdistribution hazard ratio for HIV negative patients was 1.47 *(95% CI: 1.13, 1.91)* and was significant with a p-value of 0.004. This meant that the HIV negative patient had a 47% higher subdistribution hazards to encounter discharge in the hospital than the HIV positive

patients. The hazard of being discharged from hospital decreased with age of a patient. The subdistribution hazard for age was 0.98 *(95% CI: 0.97, 0.996)* which is not so different from 0.99 *(95% CI: 0.98, 0.997)* cause-specific hazard ratio realized. Male patients, with a hazard ratio of 0.74 *(95% CI: 0.62, 0.87)* had a 26% lower subdistribution hazard of being discharged from hospital. This is not far from the cause-specific hazard ratio of 0.74 *(95% CI: 0.62, 0.88)* for the male patients.

For the event death, the results in *table 10* show that the effect sizes for the cause-specific and subdistribution hazards are fairly close again. The results indicate that the covariates interacted with the two event types but to a limited extent. *Table 4-10* shows the results for the subdistribution hazards models and cause-specific hazards models.

Table 10: Estimates of subdistribution hazards and cause-specific hazard models for death event

Model Type	**Models**	**Hazard Ratio**	**95% CI**	**P-value**
Subdistribution Hazards	*HIV Negative* (Reference: HIV Positive)	0.24	0.08, 0.79	0.018
	Age	1.03	1.004, 1.05	0.019
	Male Patients (Reference: Females)	1.36	0.75, 2.50	0.315
Cause-Specific Hazards	*HIV Negative* (Reference: HIV Positive)	0.28	0.09, 0.92	0.036
	Age	1.02	1.001, 1.05	0.042
	Male Patients (Reference: Females)	1.28	0.70, 2.36	0.426

The cause-specific hazard estimates and subdistribution hazard results were close or fairly similar by looking at the hazard ratios and coefficients. This again validates the assumption of independence of events made when applying cause-specific hazards.

From *table 10*, it is clear that the effect of patients' sex on cause-specific or subdistribution hazard were not evident enough by looking at the p-values or the corresponding confidence intervals. The patient's HIV status and age came out perfectly significant. With a subdistribution hazard ratio of 0.24 *(95% CI: 0.08, 0.79)*, HIV negative patient had by far less hazard of encountering death in hospital. Single unit older patients had a 3% higher hazard of encountering death in hospital as compared to one year younger patients. Sex was insignificant in the model with a p-value of 0.315.

On the health outcomes of the patients; 702 (84.3%) were discharged alive from the hospital, 127 (15.7%) were reported to have died in hospital while receiving medical treatment.

4.3. Model Assumptions and Goodness-of-Fit

This section presents the results for the assessment of model adequacy. The proportional hazards assumption for the Cox model was performed. Cox-Snell residual test was performed to goodness-of-fit and Martingale residual plot were used to assess function form of the covariate age.

4.3.1. Checking the Proportional Hazards Assumption of the Cause-Specific Hazards for the Event Discharged

Table 11 below presents the results obtained after carrying out proportional hazards assumption test on the three cause-specific models fitted in section 1.2.3. The failure event was patients being discharged from hospital.

Table 11: Proportional hazards assumption test for the three cause-specific hazard models

CSH Model	Chi-Square	DF	P-Value
HIV Status	8.31	1	0.004
Patient's Sex	1.99	1	0.158
Age	0.31	1	0.578

From the proportional hazards assumption test results for each model in the *table 11*, it appears that only the model with HIV status as a covariate did not meet the proportion hazards assumption. The null hypothesis was certainly rejected with the test p-value of 0.004. Theremaining two models met the proportional hazards assumption by looking at their p-values.This meant that the results from the HIV status model were not exactly accurate as the model assumptions were violated. The proportional hazards assumption for the other two models, patient's sex and patient's age, was met. This meant that for the model sex, the results were valid and accurate as the proportional hazards assumption was met. Although the age met the proportional hazards assumption too, further model assessment was done since age was fitted as a continuous variable.

The linearity of residuals for patients' age was assessed using Martingale residuals. *Figure 8* presents the output results:

Figure 8: Martingale residual plot of patients' age and event discharge.

From *figure 8*, the smoother was roughly flat and horizontal, providing no indication of the need to transform the covariate age. Therefore, with the proportional hazards assumption met, the results from the age cause-specific model were acceptable too.

4.3.2. HIV Status as a Time-varying Covariate

Since the HIV status covariate did not satisfy the proportional hazard assumption, one of the reasons could have been that HIV status was a time-varying covariate. In order to verify this, another Cox model was specified with HIV status as a time-varying covariate interacting with analysis time. The following results were yielded:

Table 12: Patient's HIV status as a time-varying covariate

Model	**Hazard Ratio**	**95% CI**	**P-Value**
Main: HIV Negative (Reference: HIV Positive)	3.05	1.71, 5.44	$P < 0.001$
Time-Varying: HIV Negative	0.81	0.70, 0.94	0.005

The estimated hazard ratios in Stata are split into two categories; those for constant-with-time variable (main) and those for time-varying covariate. From the table, the hazard ratio 0.81*(95% CI: 0.70, 0.94)* can be interpreted that the HIV negative patients had their hazard of encountering hospital discharge decreased with survival time. The patient's HIV status and survival time interacted significantly.

4.3.3. Testing the Proportional Hazards Assumption for the Cause-Specific Models of the Competing Event Death

Two sets of cause-specific models were fit in *sections 4.2.4* and *4.2.5*. The failure event was death in hospital. The first set comprised three unadjusted models and the second set was for adjusted cause-specific. Firstly, the proportion assumption tests for unadjusted models are presented followed by the test results for adjusted models. *Table 13*; show the proportional hazards assumption test for all the three unadjusted models. From this, it can be concluded that all models fully satisfied the proportional hazards assumption. With that met, it means all the interpretations made about these models were valid and statistically accurate.

Table 13: The proportional hazards assumptions test for the unadjustedcause-specific hazard (CSH) Cox models.

CSH Model	Chi-Square	DF	P-Value
HIV Status	2.90	1	0.089
Patient's Sex	0.43	1	0.513
Age	0.05	1	0.819

Table 14 presents the estimates obtained when proportional hazards assumption was tested for all the *adjusted* cause-specific models. The figures in *table 14* are from global tests only.

Table 14: Proportional hazards assumption test for the adjusted cause-specific hazard (CSH) Cox models

CSH Model	Chi-Square	DF	P-Value
HIV Status	3.22	3	0.358
Patient's Sex	3.22	3	0.358
Age	3.22	3	0.358

The p-value figures obtained indicates that the proportional hazard assumption was met. With the proportional hazards assumption met, the results from the adjusted cause-specific are statistically viable.

4.3.4. Checking the Goodness-of-Fit Using Cox-Snell Residuals Method

All the cause-specific hazard models presented earlier had to be screened for goodness-of-fit. All of them alsomet the proportional hazards assumption except for the unadjusted HIV status model with the event 'discharge from hospital' whose results were presented in *table 5*.This model was then followed up by fitting HIV status as a time-varying covariate.

Goodness-of-fit as part of the objectives of this study, it was important to establish whether the cause-specific hazard models fitted the data perfectly, or at least up to a passable level. The cumulative hazard function of the Cox-Snell residuals was obtained. Then the cumulative hazard function of the Cox-Snell residuals was plotted against Nelson-Aalen cumulative hazard function. *Figures 9* to *14* shows the graph obtained after fitting these functions:

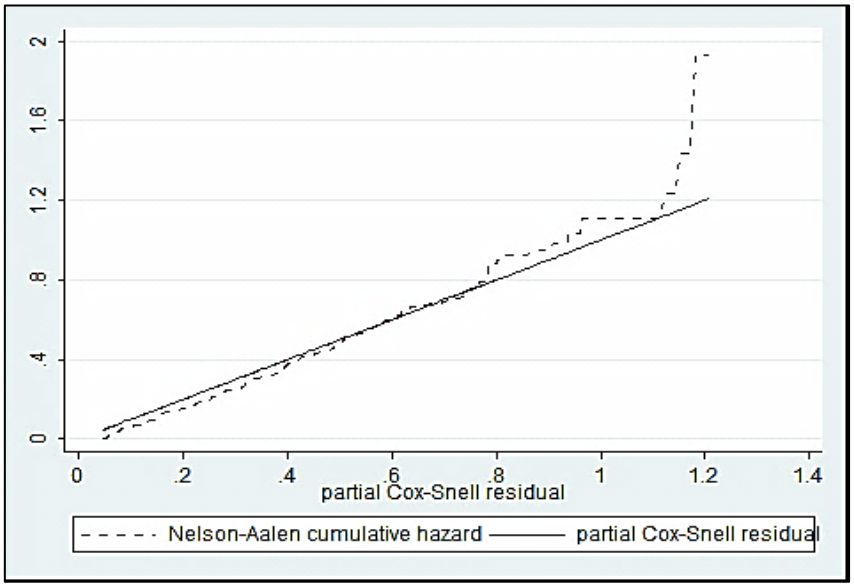

Figure 9: Cox-Snell Residual plot for patient's HIV status and event 'Discharge'

Although the proportional hazard assumption was violated by the HIV status model, the model fitted the data well by looking at how close the Cox Snell residual and Nelson-Aalen hazard curve were.

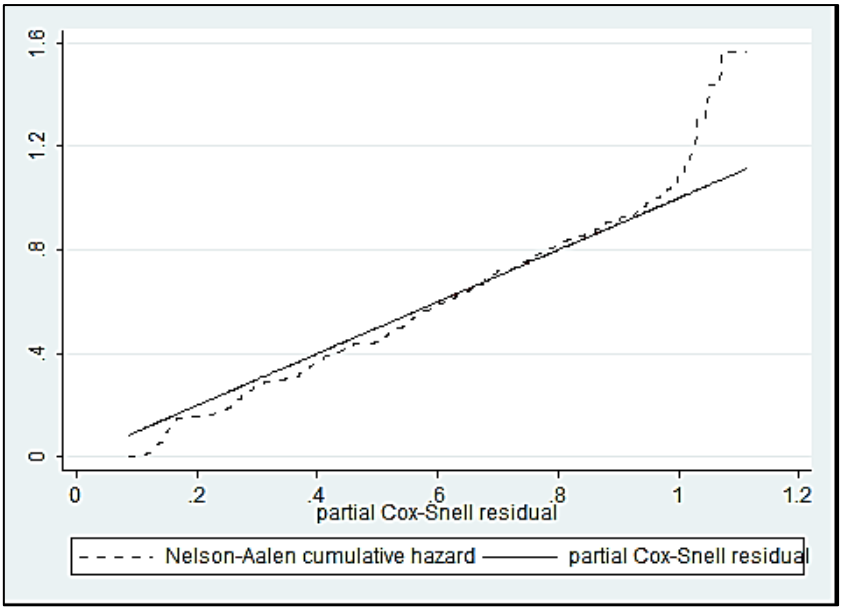

Figure 10: Cox-Snell Residual plot for patient's age and event 'Discharge'

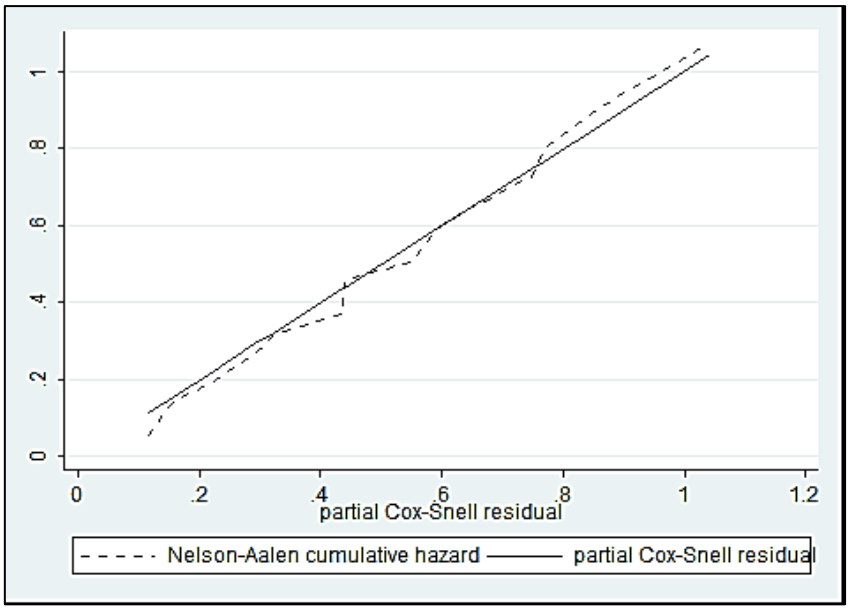

Figure 11: Cox-Snell Residual Plot for patient's Sex and event 'Discharge'

After looking and the Cox-Snell residual plots for the models with the outcome event death in hospital, further goodness-of-fit assessment was done for the competing event death.

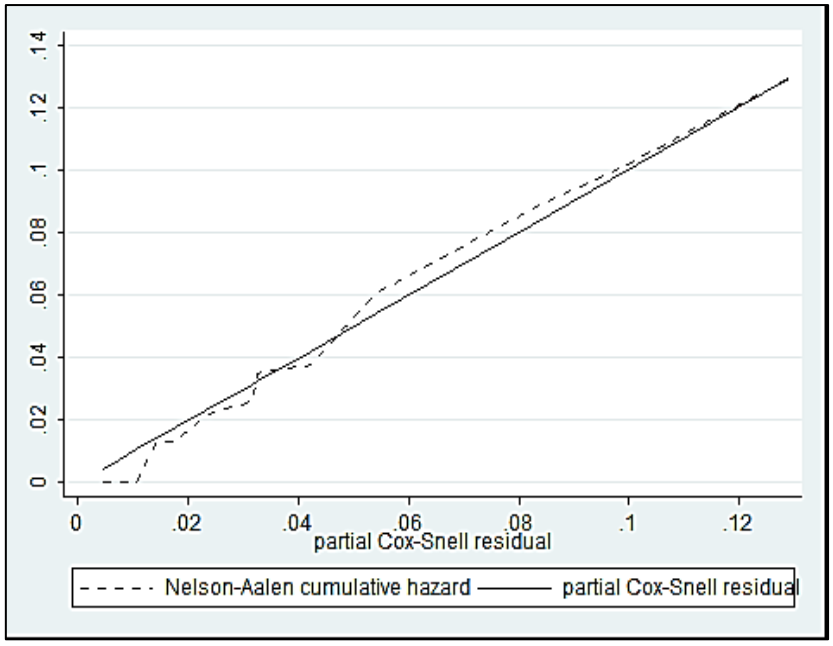

Figure 12: Cox-Snell Residual Plot for patient's HIV status and competing event 'Death'

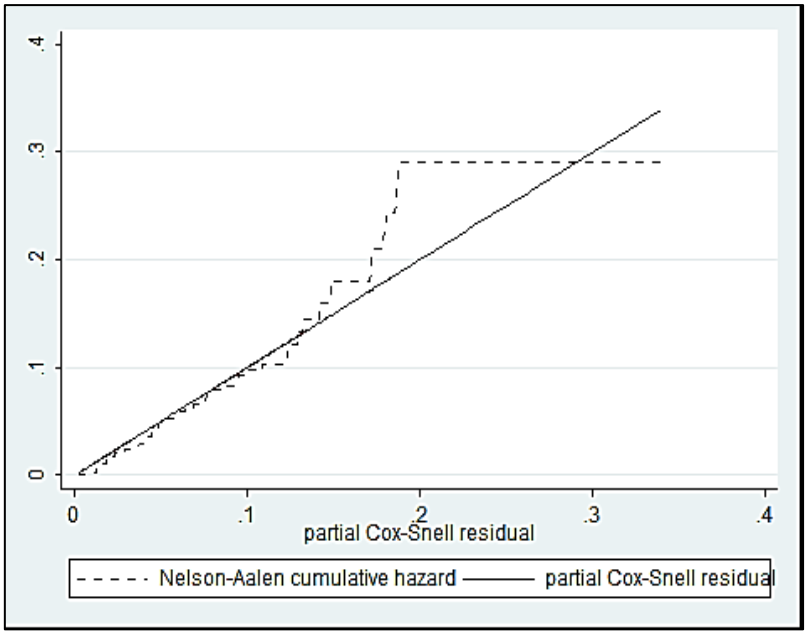

Figure 13: Cox-Snell Residual plot for patient's age and competing event 'Death'

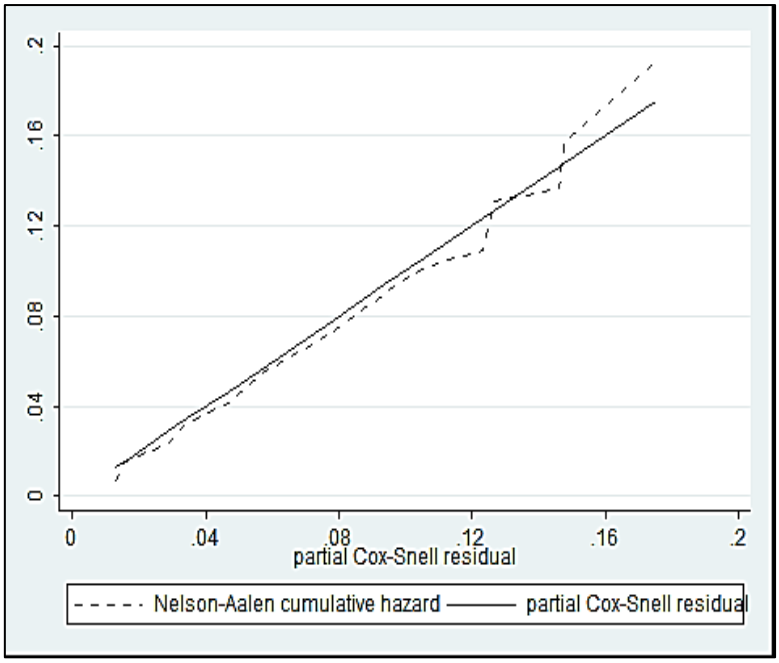

Figure 14: Cox-Snell Residual plot for patient's sex and competing event 'Death'

The Nelson-Aalen cumulative hazard curve plot was checked whether it was linear through the origin with a slope 1 as it was the case with the Cox-Snell residual function. From *figures 9* to *14*, it was observed that all models fitted the data quite well. Substantial deviations were only observed in *figure 13* where patient's age seem to encompass some outlier records. This model fit perfectly up to slightly beyond where Cox-Snell residuals were 0.15. Beyond that point; there was gross departure of the Nelson-Aalen function from Cox-Snell residual function. This meant that observations with higher values made the model not to fit the data well. In other words, the presence of high valued subjects such as the outliers in the dataset created undesirable effects on the fitted model.

4.4. Discussion of the Results

The results showed how the complement of Kaplan-Meier, $\mathbf{1} - \boldsymbol{KM}$, produced higher estimates as compared to nonparametric cumulative incidence function. At the very beginning the $\mathbf{1} - \boldsymbol{KM}$ and the cumulative incidence curve produced similar estimates. But as time went on and as some patients experienced death instead of being discharged from hospital, $\mathbf{1} - \boldsymbol{KM}$ ended up right-censoring those observation hence higher estimates. This led to $\mathbf{1} - \boldsymbol{KM}$ deviating from the nonparametric cumulative incidence function. In an event where competing risks are not present, $\mathbf{1} - \boldsymbol{KM}$ and nonparametric cumulative incidence are expected to theoretically produce same estimates. There curves are expected to superimpose. But in this case where there was death in hospital as a competing event, the best model to estimate probability that a particular event has occurred before a given time was definitely the nonparametric cumulative incidence function. Pepe and Mori test provided a comparison test for the groups' cumulative incidences. In Stata and using `pepemori` ado file written by Coviello (2004), the Pepe and Mori test automatically provides the comparison tests for both the estimated curves in the event of interest setting and for the same curves in the competing events setting without a user bothering to execute another command.

The cause-specific hazards models are best set when the assumption is that all failure events are independent. No testwas found set to specifically test this assumption. The cause-specific hazards also censors the other events not specified as failure events. A good way to contest theindependence of event assumptionwas by following up the cause-specific hazards with subdistribution hazards since the subdistribution hazards did not assume independence of events by censoring competing events. The subdistribution hazards regression is basically competing risk regression by the method of Fine and Gray (1999). The estimates obtained by the subdistribution hazards were very close to those obtained by cause-specific hazards in this

thesis. The main reason why the results from cause-specific hazard and subdistribution hazard were close was that the event of interest hospital discharge was happening by far more frequently than the competing event death. The closeness of the results between these two models can also crudely guarantee the independence of events assumption that is pre-packed with the cause-specific hazards. For this reason, the cause-specific hazards implemented in the analysis were just enough without further obtaining the subdistribution hazards estimates. However it has to be mentioned here that the choice between cause-specific hazards and subdistribution hazards greatly depends on the objectives of the study (Lau *et al*, 2009). If the interest is to see the biological mechanism of an intervention or any other covariate, the model to go for would be cause-specific hazards. If the study objective is to see the probability of an event in the presence of competing risk, then the subdistribution hazards models are better placed. For this study, the objectives centred on probability of hospital discharge, therefore the subdistribution hazards were better placed.

Testing the assumptions behind the models was done to ensure that the estimates obtained by the fitted models carried water. Ignoring the test for proportional hazards assumption can cost dearly on the creditability of the results.

Finally, testing the goodness-of-fit of the models assessed the extent to which the models fitted the observations. The Cox-Snell residuals plot was used to test on how well the model fitted the observations. The estimated Nelson-Aalen cumulative hazard function and the partial Cox-Snell residuals were obtained. The Nelson-Aalen cumulative hazard was compared to the linear residual plot with slope 1. Departures from the linear line are supposed to indicate possible lack of fit in the results. However, it has to also be mentioned that when plotting Nelson-Aalen cumulative hazard estimator for Cox-Snell residuals, even if there is a well-fitting Cox model, some variability about the 45^0 line is expected, particularly in the

right-hand tail (Machin *et al*, 2006). This is because of the effective sample caused by prior failure and censoring.

The nonparametric Cumulative Incidence Function showed that given any dayfemale patients were more likely to be discharged and less likely to die in hospital in comparison to the male patients. The unadjusted cause-specific hazard for males showed 26% less hazard of encountering hospital discharge than their female counterparts. Another notable result is that of HIV positive patients survivorship compared to the survivorship of HIV negative patients. The nonparametric Cumulative Incidence Function showed that HIV positive patients were at a greater risk of dying in hospital and lower risk of being discharge in hospital as compared to the HIV negative patients. The cause-specific hazards model estimated the risk hazard of dying in hospital for the HIV negative patients around 70% less than that of the HIV positive patients. This is noticeably a high difference and calls for further epidemiological research. There could more reasons as to why the risk difference was this big. For the cause-specific hazard, patient HIV status did not meet the proportional hazards assumption. Fitting it as time-varying explanatory variable it came out significant with a hazard ratio of 0.81. This meant that the length of hospital stay and hospital discharge interacted significantly.Other important patients' characteristics like the HIV stage and body mass index could have possibly explained further why the difference was this big. Unfortunately these characteristics were not captured in the SPINE dataset used in this thesis.

CHAPTER 5. CONCLUSION AND RECOMMENDATIONS

This chapter gives a summary of the study, outlines some of the recommendations for analysis of survival data when competing risks are present and also highlights some of the limitations of the study.

5.1. Concluding Remarks

It has been shown in this thesis how $1-KM$ produces exaggerated results as compared to the nonparametric cumulative incidence function when competing risks are available. The cumulative incidence estimates of $1-KM$ were a little larger than those obtained from fitting cumulative incidence function. This was the case because $1-KM$ treated the competing event death in the same fashion as the censored observation. However, the estimates of $1-KM$ were not awkwardly different from those of cumulative incidence function. This was so because patients were being discharged from hospital much more frequently than they were dying in hospital within the 7 day follow-up period. The nonparametric cumulative incidence function was settled for as a better estimator of survivorship since effect of death on estimating the probability of being discharged before a given day was nevertheless present.

The cause-specific hazard model estimates showed the prognostic effect of the covariates and were close to those obtained by the subdistribution hazards. The closeness of the cause-specific and subdistribution hazards was greatly attributed to the high rate occurrence of hospital discharge as compared to the competing event death. Again this meant that death in hospital had a reduced effect on the estimation of the hazard of being discharged from hospital given a follow up period of 7 days.

In an event like this where estimates from the cause-specific hazards and subdistribution hazards produces very close results; the choice of models to go for would greatly depend on the specific study objectives (Lau *et al*, 2009). As Lim *et al* (2010) put it, the advantages of cause-specific hazards is that they are more clinically understandable when assessing the prognostic effect of the covariates on a specific cause because it can be observed whether the covariate is reducing or increasing the instantaneous probability of the event of interest irrespective of other covariate effect. In this thesis, the subdistribution hazards are recommended as the interest was more on estimating the probability of being discharged from hospital given that others were dying in it. In this setting, the subdistribution hazards provided a check of the effect of the competing risk death when estimating the likelihood of being discharged in hospital. Nevertheless, it was concluded that it was important to follow up cause-specific hazards with subdistribution hazards as the subdistribution hazards might confirm or deny that the effect competing event on the estimation of the hazard of the event of interest.

The Nelson-Aalen cumulative hazard estimator for Cox-Snell residuals showed that models fitted the data well. A few deviations from the 45^0 were noted. However, even if the model indeed fit well, some departure from the 45^0 is not uncommon (Machin*et al*, 2006). This is because of the effective sample caused by prior failure and censoring.

In the analysis of competing risk data, it is important to present both the results of the event of interest and the results of competing risks. This ensures a balance of information when discussing study results and avoids overlooking the important features which may influence the interpretation of results.

Since the patients diagnosed HIV negative and the female patients were found to be more likely discharged from hospital within a short time, it is concluded that these were important

factors in determining time in hospital until discharged. The cause-specific hazards model estimated the risk hazard of dying in hospital for the HIV negative patients around 70% less than that of the HIV positive patients. This is noticeably a high difference and calls for further epidemiological research. The study also recommends further investigation on the disease specific survivorship of in-patients. In this case the interest would time to discharge for in-patients suffering from a particular disease such as malaria or tuberculosis other than an ICD-10 class of diseases.

5.2. Study Limitations

The patients' hospital survival results obtained in this study may have limited clinical use mainly because the QECH SPINE dataset did not capture other important variables such as patient's weight, height, HIV stage, literacy levels and household income which could have also possibly explained the health outcomes.

REFERENCES

Armitage, P., G, Berry, &J. N. S. Matthews. (2002). *Statistical Methods in Medical Research*(4th ed.). New York, NY: Blackwell Publishing

Brar, S. S., (2008). *Estimation of Cumulative Incidence in the Presence of Competing Risks: Application to Clinical Oncology*. (Master's Thesis, University of Calgary). Retrieved from: dspace.ucalgary.ca/bitstream/1880/46804/1/Brar_2008.pdf

Caplan R. S., Pajak T. F., & Cox J. D. (1994). Analysis of the Probability and Risk of Cause-Specific Failure.*International Journal of Radiation Oncology,* 29, 1183 – 1186.

Cleves, M., W. Gould, R. G. Gutierrez, & V. Y. Marchenko. (2010). *An Introduction to Survival Analysis Using Stata*(3rd ed.). Texas, United States: Stata Press Publication, College Station.

Cleves, A. M. (1999). Ssa13: Analysis of multiple failure-time data with Stata. *Stata technical bulletin, 4, 30 – 39.*

Coviello, V., & Boggess, M. (2004). Cumulative Incidence Estimation In the Presence of Competing Risks.*The Stata Journal, 4, 103 – 112.*

Dinse, G. E.,& Larson, M. G. (1986). A note on semi-Markov models for partiallycensored data. *Biometrika*, 73: 379–386.

Fermanian, J. D. (2003). Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis,85, 156 – 191.*

Fine, J. P., & R. J. Gray.(1999). A proportional hazards model for the subdistribution of competing risk. *Journal of American Statistical Association,* 94, 496 – 509.

Gooley, T. A., Leisenring, W., Crowley, J., & Storer, E. B. (1999). Estimation of Failure probabilities in the Presence of Competing Risks: New Representation of Old Estimators. *Statistics in Medicine Journal, 18, 695 – 706.*

Hosmer, D. W., & Lemeshow, S., (2000). *Applied Regression Analysis*.United States: John Wiley and Sons,

Kaplan, E. I., & Meier P. (1958). Nonparametric Estimation from Incomplete Observations.*Journal of American Statistics Association, 53, 457 – 481.*

Kim, H. T., (2007). Cumulative Incidence in Competing Risks Data and Competing Risks Regression Analysis.*American Association for Cancer Research*, 13(2).

Kleinbaum, D. G., & Klein, M. (2005).*Survival Analysis, A Self-Learning Text* (2nd edition). 233 Spring street, New York, N.Y.: Springer Science and Business Media, Inc.

Lau, B., Cole, S. R., & Gange, S. J., (2009). Competing Risk Regression Models for Epidemiologic Data.*American Journal of Epidemiology, 170(2), 244-256*

Lim, H. J., Zhang, X., Dyck, R., & Osgood, N., (2010). Methods of Competing Risks Analysis of End-stage Renal Disease and Mortality among People with Diabetes.*BMC Medical Research Methodology, 10, 97.*

Lin, D. Y. (1997). Nonparametric Inference for cumulative functions in competing risks studies. *Statistics in Medicine, 16, 901 – 910.*

Lin, D. Y., & L. J. Wei. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association, 8, 1074 – 1078.*

Machin, D., Cheung, Y. B., & Parmar, M. K. B. (2006). *Survival Analysis: A Practical Approach* (2nd ed). United States: John Wiley and Sons, Ltd.

Marubini E., & Valsecchi, M. G. (1998). *Analysing Survival Data from Clinical Trialsand Observational Studies.* New York, N. Y.: John Wiley and Sons, Ltd.

Nelson, W. (1982). *Applied Life Data Analysis.* Canada: John Wiley and Sons, Inc.

Pepe, M. S., &Mori, M. (1993). Kaplan-Meier, Marginal or Conditional Probability Curves in Summarizing Competing Risks Failure Time Data? *Journal of Statistics in Medicine, 12, 737 – 751.*

Pintilie, M. (2006).*Competing Risks: A Practical Perspective*.The Atrium, Southern Gate, Chichester, West Sussex, England.: John Wiley and Sons Ltd.

Satagopan J. M., Ben-Porat L., Robson, M., Kutler, D.,& Auerbach, A. D. (2004). A Note On Competing Risks in Survival Data Analysis. *British Journal of Cancer, 91(7), 1229 – 1235.*

Tai, B., Wee, J., & Machin, D., (2011). Analysis and Design of Randomised Clinical Trials Involving Competing Risk Endpoints.*Biomedical Central, 12, 127*

Vittinghoff, E., Glidden, D. V., Shiboski S. C., & MacCullochC. E.,(2005). *Regression Methods in Biostatistics.* California: Springer Science and Business Media, Inc.

Wolbers, M., Koller, M. T., Wittemam, J. C., & Steyerberg, E. W., (2009). Prognostic Models with Competing Risks: Methods and Application to Coronary Risk Prediction. Lippincott Williams and Wilkins.Epidemiology, 20