ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

SURVIVAL ANALYSISOF RECURRENT EVENTS: AN APPLICATION TO DIABETES MELLITUS PATIENTS IN THE CASE OF MENELLIK II REFERRAL HOSPITAL

Bizuwork Derebew

A Thesis submitted to

The Department of Statistics

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Statistics (Applied Statistics)

Addis Ababa University

Addis Ababa, Ethiopia

June, 2016

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAMME

COLLEGE OF COMPUTIONANAL &NATURAL SCIENCES

DEPARTMENT OF STATISTICS

SURVIVAL ANALYSISOF RECURRENT EVENTS: AN APPLICATION TO DIABETES MELLITUS PATIENTS IN THE CASE OF MENELLIK II REFERRAL HOSPITAL

Bizuwork Derebew

A Thesis submitted to

The Department of Statistics

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Statistics (Applied Statistics)

Advisor: Dr. Birhanu Teshome Ayele

Addis Ababa University

Addis Ababa, Ethiopia

June, 2016

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by Bizuwork Derbew, entitled*: Survival Analysis of Recurrent Events: An Application To Diabetes Mellitus Patients In The Case Of Menellik II Referral Hospital* and submitted in partial fulfillment the requirements for Degree on Master of Science in Statistics (Applied Statistics) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining committee:

Advisor Dr. Birhanu Teshome Signature_____ Date_____

Internal Examiner Ato. Mekonnen Tadesse Signature_____ Date_____

External Examiner Dr. Emmanuel G/Yohannes Signature _____ Date_____

Chair of Department or Graduate Program Coordinator

DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other University and all sources of materials used for the thesis have been duly acknowledged.

Name: Bizuwork Derebew

Signature: _____

Place of submission: Department of Statistics, College of Natural sciences, Addis Ababa University

Date of submission: June, 2016

This thesis has been submitted for examination with any approval as a University advisor.

Dr. Birhanu Teshome _____

Advisor's Name Signature

Abstract

Survival Analysis of recurrent events: an application to Diabetes Mellitus patients in the case of Menellik II Referral Hospital.

Bizuwork Derebew

Addis Ababa University, 2016

Diabetes mellitus is a group of metabolic diseases characterized by elevated blood glucose levels (hyperglycemia) resulting from defects in insulin discharge, insulin action or both. Recovery to normal blood sugar level in DM patients often is recurrent and correlation between events needs to be taken into account during analysis of such data. The main objectives of this research were to analysis time to recovery of DM patients and make a comparison between standard Cox-proportional Hazard (PH) and Frailty models. To achieve the objectives of the study, a retrospective data were obtained from Menellik II Referral Hospital chronic patient's clinic. All diabetes patients of over 15 years of age and who were under treatment between 2009 and 2015 were included in the study. Unmeasured shared similarities due to the impact of multiple events were modeled using a random effect (Frailty) term. The Likelihood Cross Validation (LCV) criteria were used for comparison between the Standard Cox PH and Frailty model. Shared Log-Normal Frailty model had a minimum value of LCV than the Cox PH and shared Gamma Frailty models. Hence, the shared log-normal Frailty model was chosen for analysis of the recurrent event of time to recovery of DM patients in Menellik II Referral Hospital. The median recovery time of DM patients was 32 weeks. The patient's sex and Regimen groups at baseline were significantly associated with recurrent event of time to recovery of DM patients.

Key Words: Recurrent events, Frailty, Gamma, Log-normal, LCV, Penalized Marginal Likelihood

Acknowledgement

First of all, I would like to state my limitless respect and gratitude to my **Almighty God** who sustained me to this status and for his unlimited blessing.

I would like to extend my sincere gratitude to my advisor **Dr. Birhanu Teshome** for his unreserved guidance and comments from the stage of the research proposal up to thesis write up.

My sincere acknowledgment also goes to my family, friends and colleagues who were there for me at every step of the way especially I want to say thank you to my mother; being your son is a blessing for me.

I would like to say thank Menellik II Referral Hospital manager allowing me to use their data and an employee who work in the chronic follow up clinic and the patients' documentation office for their collaboration.

Finally, my special appreciation and acknowledgment goes to Mizan-Tepi University for offering me the opportunity and financial support during my study in Addis Ababa University.

Thank you all!!!

List of Abbreviations (Acronyms)

AG-CP	Andersen-Gill Counting process
AIC	Akaike's Information Criterion
BP	Blood Pressure
Cox-PH	Cox-Proportional Hazard
DBP	Diastolic Blood Pressure
DM	Diabetes Mellitus
EM	Expectation and Maximization
FBS	Fasting Blood Sugar
GEE	Generalized Estimating Equations
HbA1c	Glycated Hemoglobin
HIV/AIDS	Human Immunodeficiency Virus /Acquired Immunodeficiency Syndrome
HR	Hazard Rate
IDDM	Insulin Dependent Diabetes Mellitus
IDF	International Diabetes Federation
LCV	Likelihood Cross-validation criterion
MICU	Medical Intensive Care Unite
M-II.RH	Menellik II Referral Hospital
MPnLE	Maximum Penalized Likelihood Estimations
MRDM	Malnutrition-Related Diabetes Mellitus
NIDDM	Non-Insulin- Dependent Diabetes Mellitus
PWP-GT	Prentice-Williams Peterson Gap Time
SBP	Systolic Blood Pressure
SSA	Sub-Saharan Africa

TB	Tuberculosis
T-I-DM	Type I Diabetes Mellitus
T-II-DM	Type II Diabetes Mellitus
WHO	World Health Organization
WLW-TT	Wei-Lin-Weissfeld Total Time

Table of contents

Contents **Pages**

List of Tables **Pages**

List of Figures **Pages**

1. Introduction

1.1. Background of the Study

Diabetes mellitus is a group of metabolic diseases characterized by elevated blood glucose levels (hyperglycemia) resulting from defects in insulin secretion, insulin action or both (Emeharole, 2008 and Grabber, *et al*., 2002). Chronic hyperglycemia is associated with micro-vascular and macro-vascular complications that can lead to visual impairment, blindness, kidney disease, nerve damage, amputations, heart disease, and stroke. In a layman's view, diabetes is termed as 'sugar ailment'. This simply means is a condition in which the body cannot change sugars and starches (carbohydrates) into energy. This happens when the body cannot make enough insulin or cannot use the insulin it makes. As a result, extra sugar in the blood can lead to damage in the blood vessels, eyes, kidneys, heart, and nerves (King *et al.,* 1998).

Diabetes is a group of diseases marked by high or low level of glucose resulting from defects in insulin production, insulin action or both. It can lead to serious complication and premature death but steps to control the disease and lower the risk of complications does exist. Insulin replacement is required for survival. The intensive-therapy regimen was designed to achieve blood glucose values as close to the normal range as possible with three or more daily insulin injections or treatment with an insulin pump. Conventional therapy consisted of one or two insulin injections per day (Leong, 2007).

Insulin is a hormone manufactured by the beta cells of the pancreas, and it is required to utilize glucose from digested food as an energy source and helps to regulate blood sugar (Emeharole, 2008).

Diabetes is divided into two main types: Type I and Type II diabetes. The former is called insulin dependent diabetes mellitus (IDDM) or juvenile-onset diabetes while the later is called adult-onset diabetes or non-insulin-dependent diabetes mellitus (NIDDM) (National Diabetes Fact Sheet United States, 2005). The main difference between Type I and II diabetes depend on cause, genetic basis, body effects, climate and diet: When Beta cells in pancreas are attacked by the body's own immune system, production of insulin will reduce leading to elevated blood glucose and causing Type I diabetes. The cause of Type II diabetes might be persistently high intake of dietary sugars that leads to excess demand on insulin production, which leads to insulin resistance over time. Genetic basis of the two main type of diabetes: in most case of Type I diabetes, the patient would need to inherit risk factors from both parents, while Type II diabetes has a stronger link to family history and lineage than Type I. Body effects, in the case of Type I diabetes are thought to be triggered by autoimmune destruction of the beta cells. Autoimmune attack may occur following a viral infection such as mumps, rubella cytomegalovirus. However, body effects in Type II appear to be related to aging, inactive lifestyle, diet, genetic influence and obesity. In addition to that, climate effect makes a difference between these Types of diabetes. Type I diabetes trigger might be related to cold weather but it develops more often in winter than in summer. When we come to Type II diabetes it is more common among people with low levels of vitamin D, which is synthesized from sunlight. Moreover, diet differ in the both types of DM. In case of Type I diabetes early diet may also play a role. Type I diabetes is less common in people who were breastfed and started solid foods at later ages. In the case of Type II diabetes, diets high in simple sugars and low in fiber and vital nutrients are more likely to lead to diabetes.

Diabetes is becoming one of the rapidly increasing non-communicable diseases and an important public health problem all over the world. Connor and Boulton (1989) notes that the main factors that lead to the cause of diabetes mellitus are hereditary (genetics) and environmental. Type-I diabetes which develops most frequently in children and adolescents can be caused by viruses that have injured the pancreas and destruction of insulin making cells by the body's immune system. Also, a family history of diabetes is a risk factor of Type-I diabetes. Type-II is a common and serious global health problem which is associated with rapid cultural and social changes, ageing populations, increasing urbanization, dietary changes, reduced physical activity and other unhealthy, lifestyle and behavioral patterns.

In 1997, an estimated 4.5% of the US population had diabetes. Direct and indirect health care expenses were estimate at $98 billion (American Diabetes Association, 2010). Diabetes is one of the serious health problems in the world and contributes directly or indirectly to thousands of deaths annually worldwide. Diabetes is an important public health challenge, it is a global non-communicable chronic disease and is largely asymptomatic, a person experiences very few signs and symptoms until damage occurs to a target organ (Hamman *et al*., 2006).

The world prevalence of diabetes among adults (aged 20–79 years) will increase to 7.7%, (439 million) by 2030. Between 2010estimate and 2030, there will be a 69% increase in the number of adults with diabetes in developing countries and a 20% increase in developed countries (IDF 5th Edition, 2011).In Ethiopia, according to WHO estimation, the number of diabetic cases in the year 2000 was 800,000 and this number is expected to increase to 1.8 million by the year 2030(WHO 2016).

In 2011, 14.7 million adults in the Africa were estimated to have diabetes, with a regional prevalence of 3.8%. The highest prevalence of diabetes in the Africa is in the island of Reunion (16.3 %), followed by Seychelles (12.4%), Botswana (11.1%) and Gabon (10.6%). Some of Africa's most populous countries also have the highest number of people with diabetes, with Nigeria having the largest number (3.0 million), followed by South Africa (1.9 million), Ethiopia (1.4 million), and Kenya (769,000). The top six countries with the highest number of people with diabetes make up over half of the total number in Africa (Diabetes Atlas 5th edition, 2011).

The prevalence of DM in Ethiopia from 1982 to 2011 was studied based on community-based studies with the estimated prevalence of DM on both types to be 4.9% among adults aged 18 years and above in the southern region (Giday and Tadesse, 2011) and 5.3% among adults 40 years and above for Type-II-DM in Oromia region (Yemane *et al.,* 2007). Hospital-based studies had found the estimated prevalence of DM to be 0.5% in all age group and 1.2% among aged 20 years and above (Blahos *et al,.* 1963, Tekelu, 1982). In addition, mortality-based studies examined mortality of patients with diabetics: among adult deaths in the capital city, 5% of them were attributed to diabetes (Mesganaw *et al.,* 2011).Study based on hospitalization of diabetes patients found the highest hospitalization in Addis Ababa (11.5%) for patients aged 60 years and above (Lester, 1982).

In this thesis, we applied survival analysis since it addresses the limitation of classical regressions like logistic and linear regressions. In survival analysis, we assess the association between survival times of events by taking into account censored observations. The Cox proportional hazard regression analysis is one of the common approach to the analysis of time to

event data. That is, conditional on the covariates, every individual has the same risk of experiencing an event such as disease recurrence (Ulviya, 2013).

Cox proportional hazard methods implicitly assume that populations are homogenous, meaning all individuals have the same risk of recovery. In the recurrent event studies, it is important to consider the population as heterogeneous, i.e. a mixture of individuals with different hazards. However, researchers often use time to first recovery of diabetic diseases from retrospective studies. These models do not take into account the fact that individuals can experience recurrent events. Recurrent events occur when subjects experience repeated occurrence of the same event. For our case, it is the time to first, second, third… recovery of patients from diabetic disease.

In this study, we applied the random effect approach, also called the Frailty models, to account for the dependence among the recurrent event times based on Andersen-Gill (A-G) survival model. The notion of Frailty provides a convenient way of introducing unobserved heterogeneity and associations into models for survival data.

Hougaard (2000) pointed out advantages of considering sources of variability into theoretically predictable and initially unpredictable even knowing relevant information separately: heterogeneity explains some `unexpected' results or gives an alternative explanation of some results, for example on proportional or decreasing hazards. The variance of the Frailty distribution determines the degree of heterogeneity in the study population.

Liang, Self, Bandeen-Roche and Zeger (1995) discussed the use of Frailty models with multivariate failure time data. The shared Frailty model was extended by Pickels *et al.* (1994) and Yashin, Vaupel and Jachine (1995) to allow different but correlated frailties among observations

within a group. Frailty models account for unobserved heterogeneity that occurs because some observations are more prone to failure, and therefore more "Frail" than others in a data set. A Frailty model is a random effects model for time-to-recovery in the DM patients' recurrent events. It has a multiplicative effect on the baseline hazard function. The overall aim of this thesis was to analyze recurrent event of time-to-recovery of diabetic patients from Menellik II Referral Hospital, (M-II-RH) using appropriate survival models.

1.2. Statement of the problem

The incidence of diabetes, especially Type II, is rapidly growing in the world. In 1985, an estimated 30 million people suffered with this chronic disease, which, by the end of 2006, had increased to 230 million, representing 6% of the world population. For a long time, Africa was considered safe from many of the diseases that are called "diseases of affluence," which plague the Western world. Similarly, there was a time when Africa was thought to be a continent relatively free of diabetes mellitus illnesses. Today, however, diabetes is very common in Africa (Diabetes, 2010).

Abraham *et al*., (2014) studied on "prediction of physical activity among type-II Diabetes patients attending Jimma University specialized hospital, Southwest Ethiopia" using logistic regression. Educational status, income per month, knowledge of the recommended physical activity, perceived severity, perceived barriers and perceived self-efficacy were significantly associated with the likelihood of engaging in the recommended physical activity. Temesgen *et al*. (2014) also studied prevalence of chronic kidney disease and associated risk factors among diabetic patients in Southern Ethiopia using logistic regression by ignoring time to event. But, such data can be more explored using survival models, such as the standard Cox PH and Frailty models, to assess the relationship between survival time and covariates on Diabetes Mellitus patients.

Inference for Cox PH model (Cox, 1972) was developed under the assumption that the observations are statistically independent and the population they come from is assumed to be homogeneous with respect to failure. However, this assumption may be violated. In many applied studies, failure times are clustered into groups such as families or geographical units. Moreover, some unmeasured characteristics shared by the members of that cluster, such as genetic

information or common environmental exposures, could influence time to the studied event. In a different context, correlated data may come from recurrent events, i.e., events that occur several times within the same subject during the period of observation. Ignoring the existence of heterogeneity will produce biased parameter estimates and inconsistent standard errors in survival analysis. According to Shepard and Zeckhauser (1977), neglecting heterogeneity overestimates the effects on life expectancy of a given medical improvement. Lancaster (1990) showed that ignoring heterogeneity causes underestimation of covariate effects in his study of unemployment rates. Henderson and Oman (1999) showed that ignoring Frailty leads to regression coefficient estimates biased towards zero by an amount depending on the distribution and the variability of the Frailty terms. For such situations, one approach for addressing the correlation is to incorporate an additive or multiplicative random effect for each cluster, resulting in a Frailty model. Random effect or Frailty model attempts to account for the existence of unmeasured attributes (such as genotype, environment and geographical location) that introduce heterogeneity into the study population. Ignoring the unobserved Frailty would underestimate model parameters. In this study, we addressed the above research problem using the Shared Frailty models.

1.3. Objectives

1.3.1. General Objective

The general objective of the study was to analyze recurrent event of time-to-recovery of diabetic patients from Menellik II Referral Hospital, Ethiopia, using appropriate survival models

1.3.2. Specific Objectives

The specific objectives of the study are:

- To fit a model that best-fits data among Standard Cox PH, shared gamma Frailty and shared log-normal Frailty models.
- To determine important factors (covariates) associated with time to recovery.
- To forward valuable recommendation that might contributes to the treatment of Diabetes Mellitus Patients and analysis of such data.

2. Literature Review

2.1. Conceptual Meaning of Diabetes Mellitus (DM)

Diabetes Mellitus

The term "diabetes mellitus" describes a metabolic disorder of multiple etiologies characterized by chronic hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The effects of diabetes mellitus include long–term damage, dysfunction, and failure of various organs (WHO, 2001).

Glucose and Diabetes

Glucose is a simple sugar, which is the body's prime source of energy. The digestive process turns the carbohydrates of a meal eaten into this glucose which is then distributed throughout the body via the bloodstream, thus, "blood sugar". The brain and other cells in the body that need immediate energy use some of the blood sugar. The rest is stored in the liver and muscles as starch called "glycogen", or in adipose tissue as "fat" to be used later. The glycogen turns back its glucose when the body needs it. A normal body maintains an even balance of sugar in the blood so as to satisfy the body's energy needs. Any disruption in this delicate balance creates a chemical imbalance either hypoglycemia-too low blood sugar level; or hyperglycemia-too high blood sugar level. Insulin, the hormone secreted by the pancreas, is what maintains the proper levels of blood sugar. However, when the pancreas fails to produce enough insulin to create a proper release of glycogen from the liver to the bloodstream, the result is high blood sugar, or diabetes mellitus.

Subjects with diabetes mellitus have blood glucose level of greater than or equal to 180mg/100ml (10mmol/l) of blood (WHO, 2011).

Symptoms of Diabetes Mellitus

The early symptoms of untreated diabetes related to elevated blood sugar levels, and loss of glucose in the urine. High amounts of glucose in the urine can cause increased urine output and lead to dehydration. In return, dehydration causes increased thirst and water consumption.

The inability of insulin to perform normally has effects on protein, fat, and carbohydrate metabolism. Insulin is an anabolic hormone, that is, one that encourages storage of fat and protein. A relative or absolute insulin deficiency eventually leads to weight loss despite an increase in appetite. Some untreated diabetes patients also complain of fatigue, nausea, and vomiting. Patients with diabetes are prone to developing infections of the bladder, skin, and vaginal areas. Fluctuations in blood glucose levels can lead to blurred vision. Extremely elevated glucose levels can lead to lethargy and coma.

2.2. Population-Based Prevalence and Incidence Studies in the World

In 2013, there were an estimated 382 million people living globally with DM. Of these 175 million people (48%) with diabetes are undiagnosed, with numbers expected to rise to 592 million by 2030. Diabetes caused 5.1 million deaths in 2013, and 80% of people with diabetes live in low-income and middle-income countries. The majority of people with diabetes were between 40 to 59 years of age (IDF 5th edition, 2011).

2.3. The Prevalence of Diabetes Mellitus (DM) in Sub-Saharan Africa

While communicable diseases such as human immunodeficiency virus/acquired immune deficiency syndrome, malaria, and tuberculosis have continued to pose greater threats to the public health system in Sub-Saharan Africa (SSA), it is now apparent that non-communicable diseases such as diabetes mellitus are undoubtedly adding to the multiple burdens that people in sub-Saharan Africa suffer. Type II diabetes mellitus (T-II-DM) is the most common form of diabetes (90–95%), exhibiting an alarming prevalence among peoples of this region. Its main risk factors include obesity, rapid urbanization, physical inactivity, ageing, nutrition transitions, and socioeconomic changes. Patients in sub-Saharan Africa also show manifestations of β-cell dysfunction and insulin resistance. However, because of strained economic resources and a poor health care system, most of the patients undergo diagnosis only after they have overt symptoms and complications. Micro-vascular complications are the most prevalent, but metabolic disorders and acute infections cause significant mortality. The high cost of treatment of T-II-DM and its comorbidities, the increasing prevalence of its risk factors, and the gaps in health care system necessitate that solutions be planned and implemented urgently. Aggressive actions and positive responses from well informed governments are desired for the conducive interplay of all forces required to curb the threat of T-II-DM in Sub-Saharan Africa. Despite the varied ethnic, transitional factors and the limited population data on T-II-DM in sub-Saharan Africa, a study by Vivian *et al (2010)* provided an extensive discussion on the epidemiology, risk factors, pathogenesis, complications, treatment, and care challenges of T-II-DM in Sub-Saharan Africa.

2.4. Prevalence of DM in Ethiopia and Literature reviews on Diabetes Mellitus

Chronic medical conditions are growing cause of death among people in developing countries. This situation is exacerbating by the migration to towns and cities of subsistence farmers and their families from the rural areas. Diabetes, though less common than in the wealthy countries such as Western Europe and North America, is not rare in Ethiopia. The condition represents a considerable health problem and is a growing cause of death in the country. The overall prevalence of diabetes in Ethiopia has been reported by various studies as follows.

Community-based studies: Two community-based studies on population prevalence of diabetes were found. A study with urban and rural sampled population in the Southern region estimated the prevalence of diabetes mellitus (Type I and II) to be 4.9% among adults aged 18 years and above (Giday *et al.,* 2011). The second study, with urban sampled population in the Oromia region, estimated the prevalence of Type II Diabetes Mellitus to be 5.3% among adults aged 40 years and above (Yemane et *al.,* 2007).

Hospital-based studies: Two studies on hospital-based prevalence of diabetes were found. In these studies, the prevalence of diabetes was estimated to be 0.5% in all age groups and 1.2% among patients aged 20 years and above (Blahos *et al.* (1963), and Tekelu, (1982)).

Mortality*:* Two studies examined mortality of patients with diabetes. Ina study, with randomly-sampled adult deaths in the capital city, 5% of deaths were attributed to diabetes (Misganaw *et al.,* 2011).

Hospitalization: Four studies investigated hospitalization of patients with diabetes, and two more studies investigated one of its complications called diabetic kitoacidosis. The highest hospitalization (11.5%) was report by a study from Addis Ababa for patients aged 60 years and

above (Lester, 1982). The prevalence for hospitalization of patients with diabetes ranged from 0.5% to 6% for aged 15 years and above (Blahos *et al.,* 1963, Lainovic, 1974). Studies estimated the prevalence of diabetic kitoacidosis to be 9.7% for patients aged 10 years and above (Bahta *et al.,* 1988) and the estimated prevalence of those admitted to Medical Intensive Care Unit was 10.7% for DM patients for all age groups (Melaku *et al.,* 2006).

According to IDF estimate, the national prevalence of diabetes among adults in Ethiopia was 4.36%, and diabetes related death among 20-79 years of age was 34,262 in 2013. Diabetes also exacerbates major infectious diseases such as TB, HIV/AIDS and malaria.

Olive *et al*. (2007) employed a frailty model to study the determinants of recovery time of diabetic patients from three hospitals in Uganda. They considered time from start of treatment to return to normal blood sugar levels as the response variable for the study and the status indicator was whether the blood sugar was in normal range or not. The independent variables included were sex, age in years, smoking status, alcoholic status, family history of the disease, body mass index, blood pressure (upper and lower), type of diabetes, duration of the disease in months, level of education, and treatment/intervention. The researchers found that treatment type known as Biguanides performs better than Insulin, diet and exercise and Sulphonylureas. Disease duration did not have a significant effect on time to remission. Time to remission was found to be inversely related to body mass index and age. Females tend to recover faster than males and the less or non-educated patients controlled the disease better than the educated ones.

Abiyot *et al*. (2014) studied time-to-first recovery of adult diabetic patients using Cox PH and shared Frailty models at Jimma university specialized hospital. They used residential places like district (Woreda) to cluster the patients and assumed unmeasured shared similarities in the cluster. They found that types of diabetics, bodyweight at baseline, fasting blood sugar level at

baseline, sex and age of patients were significantly associated with time to first recovery of diabetic patients. They also made comparison between models and shared gamma Frailty model was selected as the suitable choice for modeling time to first recovery of DM as compared to classical Cox PH model.

Giday *et al* (2014) studied risk factors associated with Type II Diabetes Mellitus in Central Zone of Tigray, North Ethiopia, and using logistic regression. They found that smoking tobacco, poor diet, physical inactivity, overweight (obesity) and hypertension were significantly associated with Type II diabetes Mellitus.

2.5. Conceptual Review about Modeling and Data Structure for Recurrent Events

The idea that an event can occur multiple times in the course of follow up of a subject is conceptually an easy extension of the single event model. We encounter things that break, are repaired, and then break again all the time. A car is a good example. Technical report by Therneau (1995) and Clayton (1994) discusses recurrent event and compares them to generalized linear models, such as Poisson and logistic regression.

A number of proportional hazard type models have been proposed for use with recurrent event data. The simplest modeling approach is to use the counting process formulation described in detail in Anderson *et al*. (1993). The approach is typically used when recurrent events are treated as identical. In this formulation, follow up is broken into segments defined by the events. If all recurrent events on the same subject are treated as identical, then the analysis required of such data is different than what is required if either recurrent events involve different disease categories and/or the order that events reoccur is considered important.

In Cox and parametric models, the hazard function may depend on unknown or non-measurable factors which can cause the regression coefficients estimated from such models to be biased. In order to overcome the problem and better model survival of patients, the Frailty models were introduced. In fact, these models are used to explain the random variation of survival function due to unknown risk factors, such as genetic factors and numerous environmental factors.

Semi-parametric inference for Frailty models was introduced by Klein *et al*. (1992), and Nielsen *et al*. (1992), as suggested by Gill (1985) they use an EM algorithm applied to the Cox partial likelihood. Hastie and Tibshirani (1993) proposed a general model with time varying coefficients and suggested estimation through penalized partial likelihood. Therneau and Grambsch (2000) noted a link between the gamma frailty model and a penalized partial likelihood. In our study, we penalize the hazard function while Therneau and Grambsch (2000) penalize the frailties.

2.6. Statistical Literature on Recurrent Events

Ullah S. *et al.,* (2014) studied on the statistical modeling of recurrent events for sport injuries and compared the following five survival models: Andersen-Gill (A-G), Frailty, Wei-Lin-Weissfeld total time (WLW-TT) marginal, Prentice-Williams Peterson gap time (PWP-GT) conditional models for the analysis of recurrent injury data. Empirical evaluation and comparison of different models were performed using model selection criteria and goodness-of-fit statistics. The Cox-PH model provided the poorest fit to the recurrent sports injury data. The fit was improved with the A-G and Frailty models compared to WLW-T T and PWP-GT models.

Issaka S. *et al*., (2014) studied malaria recurrent episodes data and compared four statistical models using simulation studies. The models were: Andersen–Gill counting process (AG-CP),

Prentice-Williams-Peterson counting process (PWP-CP), a shared gamma frailty model, and Generalized Estimating Equations model (GEE) using Poisson distribution. In that study, AG-CP and Shared gamma frailty models provided similar estimations of treatment effect but GEE Poisson distribution model failed to detect the effect of interest.

Rondeau V. *et al.,* (2010) applied statistical models for recurrent events and death of cancer patients. They applied frailty model extensions like shared frailty model: cure frailty model for a mixture of susceptible and insusceptible subjects for the event of interest, Nested frailty model when the data are clustered at several hierarchical levels, and Joint frailty model for the joint analysis of recurrent events and death using the "frailty pack" in "R package". The authors performed a semi-parametric penalized likelihood approach for parameter estimation. They focused on the different extensions of the simple shared Frailty models to analyze recurrent events not a comparison.

3. Data and Methodology

3.1. Data

3.1.1. Study Design and Data source

A retrospective six year Diabetic disease cohort study was conduct in patients enrolled to the DM follow up center. The data for this study was extracted from the total of 347 DM patient's aged greater or equal 15 years enrolled in September 2009 and then monthly followed up until August 2015. Each patient was identified with a unique medical register card code and the necessary information for our study was extracted from their follow up history (Medical Document) at Menellik II Referral Hospital, Addis Ababa.

3.2. Variables

3.2.1. **Response variable**

The outcome variable considered in this study was the time to recovery of diabetic patients. Time to recovery (might be one time, two time, or more than two times) means the time until patients come to normal fasting blood sugar level for recurrent time in the follow up period according to WHO scale (70-130 mg/dl). Right censoring was considered when patient did not recover once between the study time, transferred to another hospital, died before recovery to normal blood sugar level or before the study end.

3.2.2. **Explanatory (Independent) variables**

Covariates or independent variables used in the study:-

Sex of Patients	Employment Status
Age of patients	Educational Status
Type of DM	Regimen
Family History	Specific Drug used (treatment used)
Past medical history	Systolic Blood Pressure
Complication history	Diastolic Blood Pressure
Marital Status	Weight of Patients

Details of the variables coding are presented in Appendix (Table A1).

Clinical and Operational Definition of some covariates.

Regimen Group: DM Patients were classified into regimen groups according to the medication they took in follow up time as:-

1. Oral Agents Only**:** all are administered orally for DM patients.
2. Insulin Agents Only**:** is provided in a constant proportion to remove excess glucose from the blood, which otherwise would be toxic. It is given by injecting patients with the dose level ordered by the physician.
3. Both Oral and Insulin agents**:** is another regimen group when patients are ordered to take both oral and insulin agents at the same time in follow period.

Blood Pressure (BP) in (mmHg):- measures the pressure in blood vessels of diabetic patients. It can be classified as Systolic and Diastolic Blood Pressure.

Family History of DM: having a family who had Diabetic Disease or not.

Past Medical History: Whether the patients have other medical history before enrollment into a chronic follow up clinic as a diabetic patient.

Complication History: when the patients have other burden of disease such as Eye problem or Kidney Problem that is related to Diabetic Disease.

Note: according to the data we obtained from patients history and also from a research conducted by Olive et *al.* (2007), the Regimen group at baseline were administered for both Types of DM.

3.2.3. **Data Structure for Modelling Recurrent Event Data on DM Disease**

Diabetes Mellitus patients' data has been checked carefully to identify first, second and additional times of time to recovery on their Fast Blood Sugar Level (i.e. recurrent event data).

A sub-sample data of three patients in Table 1 illustrates data structures required for modelling recurrent event of time to recovery of diabetic disease. The time from the start of the follow-up (September 2009) to first time to recovery, second time to recovery, and the time to the last censorship in patients who did not have recovery time of diabetic disease was considered to model recurrent event data.

On the other hand, data from Diabetic disease patients with both one and more time to recovery were accounted when modelling recurrent event data. The time interval for recurrent events of time to recovery for each patient was given by the difference between two successive (consecutive) recovery times of diabetic disease patient. The starting time is the time at which a patient's FBS becomes abnormal and the ending time is the time at which a patient experiences a normal FBS (or have a normal FBS level measured in mg/dl). For instance, the follow-up history of one patient is described by the first four rows of Table 1. The first row indicates that the patient began the follow-up at time 0 and remained at risk until the 14^{th} week when an event was

confirmed; the second row indicates that the risk of another event of interest in the patient re-started 39 week after the previous event (at 53th week) and ended on 105th week when the event reoccurred; the third row indicates that the risk re-started 37 weeks after the pervious event of interest (at 142th week) and ended on 178th week; and the fourth row indicates that the patient was followed until, the 197th week and no longer reoccurrence of additional time to recovery of diabetic disease. The second patient had two time to recovery events (i.e. at 83thand 166th weeks) and the follow-up ended on the 183th week. The third patient had three time to recovery at 24th, 103th, and 203thweeks and the follow-up ended on 302th week. There were patients whose follow up history ended without an event of interest. In general, table 1 shows the follow-up history of a sample of three patients and the data structure representation for recurrent event of time-to-recovery models according to a counting process (calendar timescale).

Table 1: Data structure for modeling the recurrent time-to-event outcomes.

Id	Start time	End Time	Time	Status	sex	Regimen	Tyepe.dm
1	02/04/2002	05/19/2002	14	1	0	1	1
	0	14					
1	01/01/2003	01/03/2004	52	1	0	1	1
	53	105					
1	09/15/2004	05/29/2005	36	1	0	1	1
	142	178					
1	05/30/2005	10/12/2005	19	0	0	1	1
	178	197					
2	08/17/2004	03/23/2006	83	1	0	2	1
	0	83					
2	08/01/2006	10/29/2007	64	1	0	2	1
	102	166					
2	10/30/2007	03/08/2008	18	0	0	2	1
	166	184					
3	06/15/2002	12/03/2002	24	1	0	0	1
	0	24					
3	05/16/2003	06/14/2004	56	1	0	0	1
	47	103					
3	12/02/2004	10/09/2006	96	1	0	0	1
	127	223					
3	02/10/2007	04/18/2008	61	0	0	0	1
	241	302					

Data dictionary: ID: patients identification number; start time: the start time of the interval (in weeks); end time: the time (in weeks) at which the event occurs or the time of censoring; status, the occurrence of recovery to normal FBS (event = 1, censor = 0); time: the number of Weeks at risk that is calculated subtracting start time from end time variables; sex: sex of diabetic patients: Regimen: the regimen type that the patients used: Type.DM: the type of diabetic mellitus and others variables also exist in the main data.

3.3. Methodology

Survival Analysis

This thesis applied survival analysis to assess the survival status on time to recovery of Diabetic patients at Menellik II Referral Hospital by taking into account censoring. According to Collett (2003), survival analysis is used to describe the analysis of data that corresponds to the time from a well-defined time origin until the occurrence of some particular event or end-point. Survival data are not amenable to standard statistical procedures used in data analysis mainly due to censoring. One of the features of survival data that renders standard methods inappropriate is that survival times were frequently censored. The use of survival analysis as opposed to the use of other statistical method is most important when some subjects are lost to follow up or when the period of observation is finite (certain patients may not experience the event of interest over the study period). In this latter case, one cannot have complete information for such individuals. These incomplete observations are referred to as being censored.

In reality, such event can occur due to the following reasons:

1. A person does not experience the event before the study ends
2. A person is lost to follow-up during the study period and
3. A person withdraws from the study for unknown/ known reasons

There are three categories of censoring.

i) **Right censoring:** Survival time is said to be right censored when it is recorded from its beginning to a defined time before its end time.

ii) **Left censoring:** Survival time is said to be left censored if an individual develops an event of interest prior to the beginning of the study.

iii) **Interval censoring:** Survival time is said to be interval censored when it is only known that the event of interest occurs within an interval of time but the exact time of its occurrence is not known.

In this study, the most common type of censoring, right censoring, was used.

3.3.1 Descriptive Method of Data Analysis

Descriptive statistics are the numerical, graphical, and tabular techniques for organizing, presenting, and analyzing data. The survivor function and hazard function are the two functions of central interest in summarizing survival data. The actual survival time, *t*, of an individual is the value of a random variable time T, which can take any non-negative value. When the random variable T has a probability distribution with underlying probability density function f(t), the distribution function of T is then given by $F(t) = P(T \leq t)$ and represents the probability that the survival time is less than some value t.

Survival Function

The survivor function *S (t)* gives the probability that a person survives longer than some specified time *t*. The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values of *t* provides crucial summary information from survival data.

$$S(t) = P(T > t) = 1 - F(t), t \geq 0 \qquad (1)$$

Hazard Function

The hazard function,$\lambda(t)$gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time *t*. Thus the hazard function$\lambda(t)$ is defined as:-

$$\lambda(t) = \lim_{\Delta_t \to \infty} \frac{P[t \leq T < t + \Delta_t | T \geq t]}{\Delta_t} : t > 0 \qquad (2)$$

$$\lambda(t) = \frac{f(t)}{s(t)} = \frac{-d \log s(t)}{dt} \qquad (3)$$

Therefore, the hazard function focuses on failing, that is, on the individual will experience the event.

Relationship between Survival Function and Hazard Function:

$$s(t) = \exp\left[-\int_0^t \lambda(u)d(u)\right] = exp(-H(t)) : t \geq 0 \qquad (4)$$

where $H(t) = \int_0^t \lambda(u)d(u)$is called the Cumulative Hazard Function, which is equal to H (t) = $-\ln S(t)$.

The survival function is most useful for comparing the survival progress of two or more groups whereas the hazard function gives a more useful description of the risk of failure at any time point.

Non-parametric test: Kaplan-Meier (K-M) estimator and Log rank test

The Kaplan-Meier (KM) estimator is the standard non-parametric estimator of the survival function *S(t),* proposed by Kaplan and Meier (1958), and is also called the Product-Limit estimator. KM estimator incorporates information from all of the observations available,

both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. In addition, Log-Rank method is one of the non-parametrical statistical tests that have been proposed to answer whether there are difference of survival time between groups (Hosmer 2008). It is constructed by computing the observed and expected number of event of interest in one of the groups at each observed time to event of interest. The calculation of each test is based on a contingency table of groups by status at each observed survival time. These methods are appropriate for survival data that deal with a single (first) time to event but not for recurrent events.

3.3.2. **Standard Cox-Proportional Hazard (PH) Model**

Unlike logistic regression, Cox proportional hazard model assesses the relationship between the survival time and covariates upon the information available and taking censoring into account. The assumption of proportional hazards is that the hazard of time-to recovery at any given time for an individual in one group is proportional to the hazard at that time for an individual in the other group. Further assumptions of the Cox PH model are: (1) the ratio of the hazard function for two individuals with different sets of covariates does not depend on time; (2) time is measured on a continuous scale; (3) censoring occurs randomly; (4) uninformative censoring: standard methods used to analyze survival data with censored observations are valid only if the censoring is 'uninformative'. In practical terms, this means that censoring carries no prognostic information about subsequent survival experience.

According to Kleinbaum and Klein (2005), modeling recurrent survival data can be carried out using a Cox PH model with the data layout constructed so that each subject has a line of data corresponding to each recurrent event. The model is typically use to carry out the counting process approach is the standard Cox PH Model. For recurrent survival data, a subject remains in the risk

set for more than one time interval until his or her last interval, after which the subject is removed from the risk set. In contrast, for non-recurrent event data, each subject is removed from the risk set at the time of failure or censorship. Nevertheless, for subjects with two or more intervals, the different lines of data contributed by the same subject are treated in the analysis as if they were independent contributions from different subjects, even though there are several out-comes on the same subject. In contrast, for the standard Cox PH model approach for non-recurrent survival data, different lines of data are treated as independent because they come from different subjects. Therefore, in this study we applied Cox PH models that are used to model recurrent events.

In this study, we assumes the case of right-censored data. For the j^{th} ($j=1,\ldots,n_i$) individual in the i^{th} group ($i=1,\ldots,G$), let the T_{ij}denote the survival times under study and let C_{ij}be the corresponding right-censoring times. The observations are Y_{ij}=min (T_{ij}, C_{ij}) and the censoring indicators are$\delta_{ij} = I_{(T_{ij} \leq C_{ij})}$.The set of values of the explanatory variables in this model are represented by vector $X = \left(x_{1ij}, \ldots, x_{pij}\right)'$ which denotes the covariates for the j^{th} individual in i^{th} group. Let$\lambda_o(t)$ be the hazard function for an individual for whom the values of all explanatory variables that make up the vector X are zero. The function $\lambda_o(t)$ is called the baseline hazard function. The hazard function for the individual can then be written as:-

$$\lambda_{ij}(t) = \lambda_o(t)e^{\beta' \boldsymbol{x}} \qquad (5)$$

where β is a p x l vector of regression coefficients.

X_{ij} is the value of **x** for the j^{th} individual in i^{th} group i=1… G, and j=1… n_i

The hazard function in the Cox model is called semi-parametric since it does not explicitly describe the baseline hazard function, $\lambda_o(t)$.

3.3.2.1. Testing the Assumption of Proportional Hazards

It is always a good practice to check the assumption of proportional hazards before proceeding further with other inferential activities. Schoenfeld residuals can be used for this purpose.

Schoenfeld residuals: Schoenfeld residuals are useful to check the proportionality of the covariates over time, that is, to check the validity of the proportional hazards assumption. The expected value of the ith Schoenfeld residuals for jth explanatory variable is given by$E(r^*_{pji}) \approx \beta_j(t_i) - \widehat{\beta_J}$, where,

- $\beta_j(t_i)$ the value of time varying coefficient of X_j at the ith death time.
- $\widehat{\beta_J}$ is the estimated value of β_j in the fitted Cox Model.

Plot$r^*_{pji} + \widehat{\beta_J}$ against the recovery time. A horizontal line would suggest that the coefficient of X_j is constant and the proportional hazard assumption is satisfied. If the model fits well then the residuals are randomly distributed without any systematic pattern around the zero line, the reference line.

3.3.3 Frailty Models

Inference for Cox proportional hazards model (Cox, 1972) was developed under the assumption that the observations are statistically independent, at least conditionally upon covariates. However, this assumption may be violated. Thus in many epidemiological studies, failure times are clustered into groups such as families or geographical units: some unmeasured

characteristics shared by the members of that cluster, such as genetic information or common environmental exposures could influence time to the studied event. In a different context, correlated data may come from recurrent events, i.e. events which occur several times within the same subject during the period of observation. In frailty models, dependence is produced by sharing an unobserved variable which is treated as a random effect, or frailty (Clayton, 1978; Hougaard, 1995; Petersen, Andersen and Gill, 1996).

The term Frailty itself was introduced by Vaupel *et al*., (1979) in univariate survival models and the model was substantially promoted by its application to multivariate survival data in a seminal paper by Clayton (1978) (without using the notion "Frailty ") on chronic disease incidence in families. Frailty models are extensions of the proportional hazards model, known as the Cox model (Cox, 1972), the most popular model in survival analysis. Normally, in most clinical applications, survival analysis implicitly assumes a homogenous population to be studied. This means that all individuals sampled into that study are subject in principle under the same risk to recovery time of diabetic disease. In many applications, the study population cannot be assumed homogeneous but must be considered as a heterogeneous sample, i.e. a mixture of individuals with different hazards. For example, in many cases it is impossible to measure all relevant covariates related to the disease of interest, sometimes because of economic reasons, sometimes the importance of some covariates is still unknown. The Frailty approach is a statistical modeling concept that aims to account for heterogeneity caused by unmeasured covariates. In statistical terms, a Frailty model is a random effect model for time-to-recovery data where the random effect (the Frailty) has a multiplicative effect on the baseline hazard function.

The shared Frailty Model is given by:

$$\boldsymbol{\lambda_{ij}\left(t/X_{ij,}\,Z\right) = Z\lambda_o(t)e^{X_{ij}{}'\beta}} \qquad (\mathbf{6})$$

where Z is a Frailty effect.

To model the effects of unobserved individual differences in longevity characteristics we define Z as a measure of Frailty and assume that Z operates multiplicatively on either the total mortality hazard rate or, more generally, on a component of the total mortality hazard rate. Specification of the effects of Z in terms of the hazard rate is consistent with biological theories of the mortality process (Sacher and Trucco 1962). The multiplicative form of this specification is well known because of its use in the Cox regression model (Cox 1972; Kalbfleisch and Prentice 1980), where it is used to assess the effects of observed physiological factors on survival. Frailty models account for over-dispersion and/or clustering in survival data. They attempt to account for unobserved heterogeneity that occurs as some observations are more failure prone and hence, more "frail" than other observations in a data set. The basic idea is to introduce an additional random parameter that accounts for the random frailties. The concept of Frailty was introduced by Vaupel et *al*. (1979) who studied the model with Gamma distributed Frailties.

3.3.4. A Shared Frailty Model

A natural extension of the univariate Frailty model would be multivariate Frailty models where individuals are allowed to share the same Frailty value. The assumption of a Shared Frailty model is that both individuals in a pair share the same Frailty Z, and this is why the model is called the Shared Frailty model. It was introduced by Clayton (1978) and extensively studied in Hougaard (2000), Therneau and Grambsch (2000), Duchateau et *al*. (2002), (2003) and Duchateau and

Janssen (2004). These frailties may be individual-specific or group-specific thus giving rise to the nomenclature "individual Frailty" or "shared Frailty" models. Shared Frailty models are appropriate when you wish to model the frailties as being specific to groups of subjects, such as subjects within families. Here a Shared Frailty model may be used to model the degree of correlation within groups; i.e., the subjects within a group are correlated because they share the same common Frailty.

3.3.4.1. A Shared Gamma Frailty Model

The gamma distribution $\Gamma(k, \lambda)$has been widely applied as a frailty distribution.The two-parameter gamma density function is given by:

$$f_Z(z) = \frac{\kappa^\lambda z^{\lambda-1} \exp(-kz)}{\Gamma(\lambda)} \qquad (7)$$

with $\lambda > 0$ the shape parameter and $\kappa > 0$ the scale parameter. The Laplace transform is

$$L(s) = \int_0^\infty \exp(-zs)\, f_Z(z)\, d_z = \kappa^\lambda (s+\kappa)^{-\lambda} \qquad (8)$$

Since $L(s)$ exist in a neighborhood of zero, the mean and the variance can be obtained by using the first and second derivatives of the Laplace transform

$$L^{(1)}(s) = -\lambda \kappa^\lambda (s+\kappa)^{-\lambda-1} \qquad (9)$$

$$L^{(2)}(s) = -\lambda(\lambda+1)\kappa^\lambda (s+\kappa)^{-\lambda-2} \qquad (10)$$

Evaluating these derivatives at s=0 we find

$$E(Z) = (-1)L^{(1)}(0) = {}^{\lambda}/_{\kappa} \qquad (11)$$

$$Var(Z) = L^{(2)}(0) - \left(L^{(2)}(0)\right)^2 = {}^{\lambda}/_{\kappa^2} \qquad (12)$$

In Frailty modelling the typical choice of the parameters of the gamma distribution is k=λ. Using θ as notation for the variance of Z, we the have $E(Z) = 1$ and $Var(Z) = \theta - {}^{1}/_{\kappa}$. This distribution with parameters $({}^{1}/_{\theta}, {}^{1}/_{\theta})$ is called a one-parameter gamma distribution with variance parameter θ.

The gamma distribution is a flexible distribution that takes a variety of shapes as k (scale parameter) varies: when k=1, it is identical to the well-known exponential distribution; when k is large, it takes a bell-shaped form reminiscent of a normal distribution. Despite these advantages, it is necessary to mention that no biological reason exists that makes the gamma distribution more preferable than other Frailty distributions. Nearly all arguments in favor of the gamma distribution are based on mathematical and computational aspects. In contrast, a paper by Abbring and van den Berg (2007) rationalizes the use of gamma distributions for frailties in time-to-event data analysis. The authors show that, in some classes of Frailty models, the distribution of the Frailty among survivors converges to a gamma distribution under mild regularity assumptions as time goes to infinity.

With the assumption k= $\boldsymbol{\lambda}$ (necessary for identifyability reasons), the two-parameter gamma distribution turns to a one parameter distribution$\boldsymbol{\Gamma\left(\frac{1}{\theta},\ \frac{1}{\theta}\right)}$. A mathematically convenient choice for the distribution of Z_i and one that is considered in our study is the one parameter gamma distribution. The functional form of the one parameter gamma distribution is given by:

$$f_z(z) = \frac{Z^{\frac{1}{\theta}-1}\exp(-Z/\theta)}{\theta^{1/\theta}\,\Gamma\,(1/\theta)}\ ,\theta > 0 \qquad (13)$$

where $\Gamma(.)$ is gamma function with Laplace transform that is given by:

$$L(s) = (1+\theta s)^{-1/\theta}, \theta > 0 \qquad (14)$$

Thus, the expectation and variance of the Frailty variable will be 1 and θrespectively.Due to the simplicity of the Laplace transform, it is easy to derive closed form expressions of the conditional survival, cumulative density, and hazard function. This is also the reason why Gamma distribution has been used in most applications. In this situation, individuals j in a cluster i are supposed to share the same Frailty, Z_i. The unconditional survival function of the gamma Frailty distribution is given by:-

$$S_\theta(t) = [(1-\theta ln\{S(t)\})]^{-1/\theta}, \quad \theta > 0 \qquad (15)$$

The unconditional hazard function of the gamma Frailty distribution is given by:

$$\lambda_\theta(t) = \lambda(t)[1-\theta ln\{S(t)\}]^{-1} \qquad (16)$$

Where *S (t)* and *λ (t)* are the survival and hazard function in the distributions.

The shared gamma Frailty model (conditional hazard) for individual j in cluster i is:

$$\boldsymbol{\lambda_{ij}(t|X_{ij},Z_i) = Z_i\lambda_o(t)e^{X_{ij}'\beta} = Z_i h(t_{ij})} \qquad (\mathbf{17})$$

where $h(t_{ij}) = \lambda_o(t)e^{X_{ij}'\beta}$in the Cox regression model for individual j in cluster i. The Z_i are independent identically distributed following a gamma distribution, like in the univariate

Frailty models. This model is therefore an extension of the described model. The interpretation of this model is that the between-groups variability (the random variation of Z) leads to different risks for the groups, which then show up as dependence within the group.

3.3.4.2. Shared Log-Normal Frailty Model

Another important Frailty distribution is the log-normal distribution. The popularity of the log-normal Frailty model stems mainly from the link with mixed models, where the standard assumption is that the random effects follow a normal distribution. Let W∼N (0,σ^2) be a normally distributed random effect and let the frailty be given by $Z = e^W$. The corresponding Frailty has a lognormal distribution. It is function has the form:-

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\sigma^2}}\, e\left(\frac{-(logz)^2}{2\sigma^2}\right), \sigma^2 > 0, \qquad Z > 0 \tag{18}$$

with $\sigma^2 > 0$. In this case,the expectation and variance of the Frailty Z are functions of the parameter σ^2.The expected value and variance of Z are given by:-

$$E(Z) = e^{\frac{\sigma^2}{2}} \tag{19}$$

$$Var(Z) = e^{\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{20}$$

Consequently, in the Gamma Frailty model the parameter θdenotes the variance of the Frailty Z whereas in the log-normal model σ^2 denotes the variance of the random effect W=ln (Z). Both expressions can't be directly compared. Furthermore, in the log-normal model the expectation of the Frailty variable is usually not one despite the fact that the expectation of the random effect W is zero. The shared log-normal Frailty model has the form:-

$$\lambda_{ij}(t) = \lambda_o(t)e^{X_{ij}{}'\beta + W_i} \tag{21}$$

where $\lambda_{ij}(t)$is the hazard function for the j[th] individual from the i[th] group, $\lambda_o(t)$ is the baseline hazard at time *t*, X_{ij} is the vector of p covariates recorded for the individual and W_i is the random effect for the i[th] group. In this model $\lambda_o(t)$ can be left arbitrary. The wi's, i =1,..., G are a sample (independent and identically distributed) from a density $f_W(.)$. The frailty model can be rewritten as follows:

$$\lambda_{ij}(t) = \lambda_o(t)exp(W_i)e^{X_{ij}'\beta} = \lambda_o(t)Z_i e^{X_{ij}'\beta} \qquad (22)$$

where $Z_i = exp(W_i)$ is known as the Frailty. Model (22) is a conditional hazard function given the independent Zi's, i =1,..., G which are assumed to have a common density $f_Z(.)$.

Two classical choices for the density of the frailties are:

(a) The zero-mean normal density for W: in this case the density of Z is log-normal, i.e.

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\,\sigma^2}}\, e\left(\frac{-(logz)^2}{2\,\sigma^2}\right),\ \sigma^2 > 0, \qquad z > 0 \qquad (23)$$

$$with\ mean\ E(Z) = e^{\frac{\sigma^2}{2}}\ and\ variance\ Var(Z) = e^{\sigma^2}(e^{\sigma^2} - 1)$$

(b) The one-parameter gamma density for Z,

$$f_Z(z) = \frac{z^{\frac{1}{\theta}-1}\exp(-z/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)} \qquad (24)$$

Then the corresponding density for W is

$$f_W(w) = \frac{(\exp(w)^{\frac{1}{\theta}}\exp(\frac{-\exp(w)}{\theta})}{\frac{1}{\theta}\Gamma(1/\theta)} \qquad (25)$$

which is the log-gamma density. We note that $E(W) = \psi(\frac{1}{\theta})$ and $Var(W) = \psi^{(1)}(\frac{1}{\theta})$ where $\psi(.)$ and $\psi^{(1)}(.)$are the digamma and trigamma functions respectively.

Since Z in (22) can be thought of as a mixing term, its density $f_Z(.)$ is also referred to as a mixing distribution. Typically $Var(W) = \sigma^2$ is used to describe the heterogeneity among the groups in the log-normal density case whereas $Var(Z) = \theta$ is used in the gamma density case.

The gamma distribution has been used extensively due to its mathematical convenience that results from the simple form of its Laplace transform. This leads to closed form expressions for the unconditional (marginal) survival and hazard functions. No closed form expression exists for the Laplace transform for the log-normal distribution. On the other hand, this latter distribution is more flexible than the gamma in creating correlated frailties, thus resulting in its use in multivariate frailty models.

3.3.5. Penalized Likelihood Approach

Semiparametric hazard models without Frailty terms are fitted by maximization of the partial likelihood (Cox, 1972). For semiparametric Frailty models, however, we need to account for the contribution of the unobserved Frailty terms. Assuming a parametric baseline hazard is restrictive in the sense that detected lack of fit does not ensure that this is a direct consequence of the misspecification of the Frailty density. It would be interesting to have a method that still uses the family of Frailties distribution but leaves the baseline unspecified. For parametric Frailty

models, maximization of the marginal likelihood then leads to estimates of the parameters in the model. However, for semiparametric Frailty models, more complex estimation techniques are needed. Therefore, in this study we focus mainly on semiparametric Frailty models, which means that traditional maximum likelihood estimations procedures are not appropriate for parameter estimation. An appropriate estimation method could be used to fit semiparametric Frailty models that is the expectation-maximization (EM) algorithm and the penalized likelihood approach as discussed in Therneau and Grambsch (2000) and Duchateau and Janssen (2008). In this study, we use the panelized likelihood approach because the idea of cox partial likelihood does not carry out in a simple manner, since the integration over Frailty induces a complicated form of this likelihood. In the gamma Frailty model, a compact formula for the full likelihood can be obtained by integrating out the Frailty Z_i from the join likelihood (Klein *et al.,* (1992); Nielsen *et al.,* (1992)). Thus, the penalized likelihood approach has the advantage that while making no parametric assumption on the hazard or intensity functions, it yields smooth estimates of these functions. Moreover, we can avoid using gamma functions in the expression of the log-likelihood, thus making it easier to compute the unknown parameters. For instance, in the shared gamma Frailty models, the full log-likelihood for right censored data takes a simple form with an analytical solution for the integrals on the Frailty term.

$$l(\lambda_o(.),\beta,\theta) = \sum_{i=1}^{G}\left\{\sum_{j=1}^{ni}\delta_{ij}\{\beta' X_{ij} + \ln(\lambda_o(Y_{ij}))\} - \left(\frac{1}{\theta}+m_i\right) ln\left[1+\theta\sum_{j=1}^{ni}\Lambda_0(Y_{ij})e^{(\beta' X_{ij})}\right]\right.$$
$$+\frac{1}{\theta}\ln\left(1+\theta\sum_{j=1}^{ni}\Lambda_0(\mathcal{L}_{ij})e^{(\beta' X_{ij})}\right)$$
$$\left.+\mathrm{I}_{mi\neq 0}\sum_{k=1}^{mi}(\ln(1+\theta(mi-k)))\right\} \qquad (26)$$

with $\Lambda_0(.)$is the cumulative baseline hazard function and $mi = \sum_{j=1}^{Ji} I_{(\delta_{ij=1})}$ is the number of observed events in the i[th] group.

We introduced the semi-parametric penalized likelihood approach to jointly estimate the parameters $\beta,\ \theta,\ \sigma^2$and the baseline hazard function $\lambda_0(t)$, which is assumed to be smooth. A possible means for introducing such an *a priori* knowledge is to penalize the likelihood by a term which has large values for rough functions (i.e. complex functional form) (O'Sullivan, 1988; Joly *et al.,* 1998). Thus, we define the Maximum Penalized Likelihood Estimators (MPnLE) of $\lambda_0(t)$, β and θ as maximizing:

$$pl(\lambda_o(.), \beta, \theta) = l(\lambda_o(.), \beta, \theta) - K\int_0^{\infty} \lambda_o''^2(t)dt \qquad (27)$$

Penalized Likelihood approach for shared log-normal Frailty

Let W_i represent random effects which follow a normal distribution with zero mean andvarianceσ^2. The hazard at time t_{ij} for an individual (with shared frailty = Zi) is of the form given by model (22).

We introduce a semi-parametric penalized likelihood approach to estimate the different parameters: the regression coefficients, the variance of the random effects and the baseline hazard function $\lambda_o(t)$(Rondeau V., *et.al*. 2003). In most situation it is reasonable to expect a smooth baseline hazard function, the piecewise constant modeling for the hazard function being often unrealistic. To introduce such a priori knowledge, we penalized the likelihood by a term which has large values for rough functions (Sullivan O., (1988), and Joly P., *et. al.* 1998).Thus, we define the Maximum Penalized Likelihood Estimators (MPnLE) of $\lambda_0(t)$, β and σ^2 as maximizing

$$pl(\lambda_o(.), \beta, \sigma^2) = l(\lambda_o(.), \beta, \sigma^2) - K\int_0^{\infty} \lambda_o''^2(t)dt \qquad (28)$$

The roughness penalty function is represented by the squared norm of the second derivative of the hazard function (Sullivan O., 1988).The estimator $\widehat{\lambda}(.)$ cannot be calculated explicitly but can be approximated on a basis of splines. Splines are piecewise polynomial functions that are combined linearly to approximate a function on an interval. We used cubic M-splines, which area variant of cubic B-splines (Ramsay J., 1988).The estimated parameters are obtained by the robust Marquardt algorithm (Marquardt D., 1963) which is a combination between a Newton–Raphson algorithm and a steepest descent algorithm.

In equation (27), $l(\lambda_o(.), \beta, \theta)$is the full log-likelihood and $k \geq 0$ is positive smoothing parameter which controls the trade-off between the data fit and the smoothness of the functions. In practice, the range of the integral is restricted to the period when at least one subject is still at risk. This expression represents a trade-off between faithfulness to the data, as represented by *l* (.), and "smoothness" of the solution, as represented by the squared norm of the second derivative. For large *k* the term $\int_0^\infty \lambda_o''^2(t)dt$ will be forced toward zero and the curves $\widehat{\lambda_o}(.)$ will approach linear functions of time. If *k* is small, then the main contribution to *pl* (.) will be the log-likelihood $l(\lambda_o(.), \beta, \theta)$and the curve estimate λ_0 will track the data closely, but will be more irregular.

3.3.5.1 Choice of Smoothing Parameter

Sometimes it is sufficient to choose the smoothing parameter heuristically, by plotting several curves and choosing the one that seems most realistic. An empirical estimate of the smoothing parameter can be provided or the smoothing parameter can be chosen by maximizing cross-validation as in Joy *et al*., (1999).The cross-validation procedure was implemented for classical cox proportional hazard model and shared Frailty model.

In each case smoothing parameters could be chosen by maximizing an approximate cross-validation score which was detailed by O'Sullivan (1988) for a Cox model:

$$\overline{CV(k)} = \frac{1}{n} l_j(\hat{\eta}(k)) - \frac{1}{n}\left(\left[\hat{I}(\hat{\eta}) + 2k\Omega\right]^{-1}\hat{I}(\hat{\eta})\right) \qquad (29)$$

where:

l_j is the log-likelihood contribution of individual j.$I(\eta) = E(-\frac{\partial^2 l}{\partial \eta^2})$ is the information matrix.

$\widehat{H}(\hat{\eta}) = \hat{I}(\hat{\eta}) + 2\text{k}\Omega$ is the negative of the converged hessian of the penalized log − likelihood and $\Omega = \int \left(\frac{\partial^2 M}{\partial u^2}(u)\right)' \frac{\partial^2 M}{\partial u^2}(u)du$.

As in Gray (1992), if we interpret trace $\left(\left[\hat{I}(\hat{\eta}) + 2\text{k}\Omega\right]^{-1}\hat{I}(\hat{\eta})\right)$ as an effective number of parameters or as the model degrees of freedom, Likelihood cross-validation criterion (LCV) is equivalent to AIC criterion (Gray R., 1992). In this study, the goodness of fit of the Cox and Frailty models is provided by an approximate likelihood cross-validation criterion (LCV) (Gray R., 1992). In the case of parametric approach, LCV is approximately equivalent to Akaike's criterion. Lower values of LCV indicate a better fitting model.

3.3.5.2 Computational Procedure (Algorithm)

The estimated parameters for the models we employed were obtained by the robust Marquardt algorithm (Marquardt, 1963) which is a combination of the Newton-Raphson algorithm and steepest descent algorithms. It is more stable than the Newton-Raphson algorithm but

preserves its fast convergence property near the maximum. The iteration stops when the difference between two consecutive log-likelihoods is small, the coefficients are stable and the gradient is small enough.

3.3.6. Assessing Model Adequacy

Regardless of which type of model is fitted and how the variables are selected to be in the model, it is important to evaluate how well the model fits the data. A survival model is adequate if it represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as goodness of fit. Residuals are a useful method for checking the fit of a statistical model. Residuals are central to the evaluation of model adequacy in any setting. Cox-Snell residual is the most widely used residual in the analysis of survival data (Cox and Snell (1968)). The Cox-Snell residual for the ith individual is given by:

$$r_{ci} = \exp(\beta' X_{ij})\, \hat{H}_o(t_i) = \hat{H}_i(t_i) = -\log\left(\hat{S}_i(t_i)\right) \qquad (30)$$

where $\hat{H}_o(t_i)$is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of that individual, $\hat{H}_i(t_i)$and $\hat{S}_i(t_i)$are the estimated values of the cumulative hazard and survivor functions of the ith individual at t_i. The hazard function follows approximately 45 degree line at which plot depicts for -log$\widehat{(S_i}(t_i))$ vs t a straight line and/or for plot log [-log $\widehat{(S_i}(t_i))$] vs log (t) a straight line through the origin with slope =1, we can say the model fit the data better.

4. Result and Discussion

4.1 Descriptive survival analysis of diabetic mellitus patients

In this study, 347 diabetic patients who followed diabetic treatment in Menellik II Referral Hospital between 1st September 2009 and 30th August 2015 were considered. The outcome was time to recovery from diabetic's disease as a recurrent event data. Descriptive statistics of baseline covariates are illustrated in Table 2. From the total of 1273 events, 923(72.56%) events were recorded and the rest 350(27.44%) were censored.

Majority of the cases, 273(78.7%) out of 347 patients, were Type-II diabetic, 166(47.8%) were females, 189(54.5%) had no past medical history and 218(62.8%) had no family history of the disease. The results also showed that 199(57.3%) had a complication such as hypertension or other complication due to diabetic disease; 186(68.1%) were married. Also, the result shows that 190(54%) were educated; and 143(41.2%) were unemployed. In addition to that, 131(37.8%) were classified in Insulin regimen; 125(36.0%) of the patients took Monotend drug. Moreover, 161(46.4%) of the patients had high (>=130 mmHg) systolic blood pressure; 175(50.4%) had also high (>=80 mmHg) diastolic blood Pressure and 142(40.9%) were classified as Oral Type drug users.

Table2: The Number of DM Patients with Each Type of DM

No.	Variable Names	Category	Type of Diabetic Mellitus Disease		
			Type I DM	Type II DM	Total
1	Sex of Patents	Female	33(9.5%)	166(47.8%)	199(57.3%)
		Male	41(11.8%)	107(30.8%)	148(42.7%)
		Total	74(21.3%)	273(78.7%)	347(100.0%)
2	Past Medical History	Yes	12(3.5%)	84(24.2%)	96(27.7%)
		No	62(17.9%)	189(54.5%)	251(72.3%)
		Total	74(21.3%)	273(78.7%)	347(100.0%)
3	Family History	Yes	9(2.6%)	55(15.9%)	64(18.5%)
		No	65(18.7%)	218(62.8%)	283(81.5%)
		Total	74(21.3%)	273(78.7%)	347(100.0%)
4	Is there Complication	Yes	34(9.8%)	199(57.3%)	233(67.1%)
		No	40(11.5%)	74(21.3%)	114(32.9%)
		Total	74(21.3%)	273(78.7%)	347(100%)
5	Marital Status of Patients	Married	25(33.8%)	186(68.1%)	211(60.8%)
		Single	49(66.2%)	87(31.9%)	136(39.2%)
		Total			
6	Educational Status of Patients	educated	72(20.7%)	190(54.8%)	262(75.5%)
		uneducated	2(0.6%)	83(23.9%)	85(24.5%)
		Total	74(21.3%)	273(78.7%)	347(100%)
7	Employee Status of Patients	employee	40(11.5%)	130(37.5%)	170(49.0%)
		unemployed	34(9.8%)	143(41.2%)	177(51.0%)
		Total	74(21.3%)	273(78.7%)	347(100%)
8	Regimen	Oral Agents	2(0.6%)	91(26.2%)	93(26.8%)
		Insulin Only	68(19.6%)	131(37.8%)	199(57.3%)
		Insulin and Oral Agents	4(1.2%)	51(14.7%)	55(15.9%)
		Total	74(21.3%)	273(78.7%)	347(100%)
9	Specific Type of Drugs at Time	Doanied	3(0.9%)	83(23.9%)	86(24.8%)
		HCT	2(0.6%)	0(0.0%)	2(0.6%)
		Metformin	1(0.3%)	3(0.9%)	4(1.2%)
		Monotend	61(17.6%)	125(36.0%)	186(53.6%)
		Lute	1(0.3%)	4(1.2%)	5(1.4%)
		Regular	5(1.4%)	2(0.6%)	7(2.0%)

		all Oral	1(0.3%)	56(16.1%)	57(16.4%)
		Total	74(21.3%)	273(78.7%)	347(100%)
10	Systolic Blood Pressure at diagnosis time in mmHg	<=110(Below)	37(10.7%)	27(7.8%)	64(18.4%)
		110-130(Normal)	25(7.2%)	85(24.5%)	110(31.7%)
		>=130(High)	12(3.5%)	161(46.4%)	173(49.9%)
		Total	74(21.3%)	273(78.7%)	347(100%)
11	Diastolic Blood Pressure at each time mmHg	<=60(Below)	4(1.2%)	2(0.6%)	6(1.7%)
		60-80(Normal)	37(10.7%)	96(27.7%)	133(38.3%)
		>=80(High)	33(9.5%)	175(50.4%)	208(59.9%)
		Total	74(21.3%)	273(78.7%)	347(100%)
		Total	74(21.3%)	273(78.7%)	347(100%)

From Table 3 (descriptive statistics about continuous variables), the mean age of the patients at baseline was 49 years with the minimum and maximum age of 15 and 82 years, respectively. The mean systolic blood pressure and diastolic blood pressure of the patients at baseline were 127.28 mmHg and 77.83 mmHg respectively. The minimum and maximum Systolic Blood Pressure and Diastolic Blood Pressure of the patients were 70 and 200 mmHg and 48 and 140 mmHg, respectively. The results also show that the average weight of the diabetic patients at diagnosis time was 64.92 Kg with the minimum and maximum weight of 40 and 103 Kg, respectively. The mean and median time to recovery were 51 and 32 weeks, respectively, while the minimum and maximum recovery times were 1 and 318 weeks, respectively.

Table 3: Summary Statistics for Continuous Variable

No	Variable	Number	Range	Minimum	Maximum	Mean	Std. Error.
1	Patients Age at Baseline Time	347	67	15	82	49	.80
2	Systolic Blood Pressure at each time in mmHg in Continuous	347	130	70	200	127.28	1.24
3	Diastolic Blood Pressure at each time mmHg in Continuous	347	92	48	140	77.83	.62
4	Wight of Patients at time in K/G	347	63.00	40.00	103.00	64.92	.61
5	Summary about recurrent events of time to recovery of DM patients	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
		1	17	32	51	64	318
		n events= 923 n groups= 347					

4.2 Result: From Standard Cox-PH Model

4.2.1 Modeling Time to Recovery of Diabetes Mellitus Patients

To determine the biological, clinical and socio demography covariates which are associated with the observed time to recovery of diabetic patients, we first fitted Cox proportional hazard model for each potential risk factor before proceeding to more complicated models. Result of the univariable Cox proportional hazard regression model is shown in the Appendix (Table A2). Variable with p-value less than or equal to 0.25 in the univariable analysis were considered for multivariable model. Then, the full multivariable Cox proportional hazard model was fitted including all the potential covariates that were significant at 25% at the univariate level. For multivariable analysis forward selection method was used and variables with P-value less than or equal to 5% were selected as significant covariates.

The result of the standard Cox PH model is presented on Table 4. It is observed that sex and regimen groups were significantly associated with time to recovery. The Standard Cox PH model considers different line of data contributed by the same subjects as independent contributions from different subjects.

Table 4: Parameter Estimates of Standard Cox PH Model

Variable Names	Category	coef	exp(coef)	SE coef (H)	SE coef (HIH)	Z	P_value	95% CI.
Sex	Male	0.251	1.286	0.0675	0.0675	3.72	0.0002	(1.13, 1.47)
Regimen group	Insulin and Oral Agents	-0.262	0.769	0.0972	0.0972	-2.70	0.007	(0.64, 0.93)
	Oral Agents	-0.222	0.801	0.0818	0.0818	-2.72	0.007	(0.68, 0.94)
	Regimen		Chisq=12.0486		Df=2	global p=0.0024		
Penalized marginal log-likelihood = -4819.2 Convergence criteria: parameters = 0.00013 likelihood = 0.00027 gradient = 6.18e-09 LCV = the approximate likelihood cross-validation criterion in the semi parametrical case=3.80041								
Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively.								

4.2.1.1. Assessing Proportionality Assumption

The PH assumption of all variables included in the model was checked using the Schoenfeld residuals as described in Table 4.1. The results show that the covariates are not statistically significant implying that the covariates are time independent. The overall proportionality test is also not statistically significant implying that the proportionality assumption was not violated.

Table 4.1.: Cox PH assumption checking test statistics based on Schoenfeld residuals for DM.

Variable	Category	Rho	Chisq.	P-value
Sex	Male	0.0314	1.032	0.310
Regimen	Insulin and Oral Agents	-0.0140	0.222	0.638
	Oral Agent	-0.0242	0.618	0.432
	GLOBAL	NA	1.920	0.589

Figure 1 depicts the plot of Schoenfeld residuals for each covariate in the model against time. There are no eminent departures from proportionality. Hence, PH assumption among sex and Regimen group at baseline variables are assumed to be fulfilled.

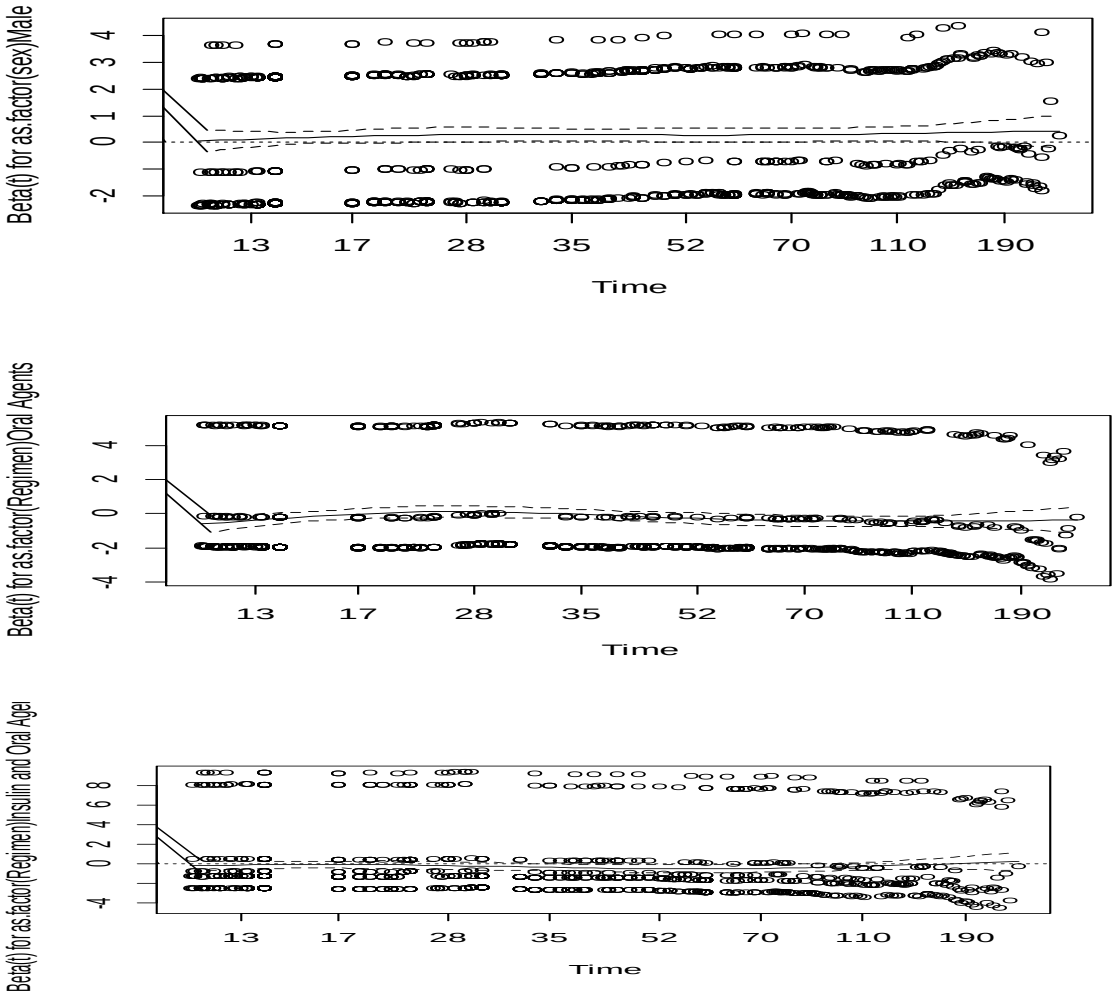

Figure 1: Plots of Scaled Schoenfeld Residuals for each Covariate in the model.

4.2.2 Shared Gamma Frailty Model

In recurrent event data, subjects may have two or more events of interest. Thus, patients with the same id are considered as correlated. An extension of the Cox model can be considered by taking into account the hierarchical (clustered) structure of the data. Thus clustering can be

considering as a random effect. Here the main interest is rather in the heterogeneity between subjects.

In the shared gamma Frailty model, first univariable analysis was conducted and variables significant at 25% level of significance were taken to the multiple shared gamma Frailty model. Result is presented in Table A3 of the Appendix. Sex and Regimen group were significant covariates selected from the saturated multiple shared gamma Frailty model.

In our study, the model based and empirically corrected standard errors ($\widehat{H}^{-1}$ and $\widehat{H}^{-1}I\widehat{H}^{-1}$) were almost identical for the Cox PH model. Comparing estimated standard errors presented in Table 4 and Table 5, slight differences were observed in the standard error from Cox PH model and shared gamma Frailty model. As expected standard errors of the Shared Gamma Frailty model were slightly larger than standard error of the Cox PH model.

Table 5: Parameter Estimates of Shared Gamma Frailty Model.

Variable Names	Category	Coef	exp (coef)	SE coef (HIH)	z	P_Value	95%CI for Hazard
Sex	Male	0.255	1.291	0.0709	3.598	0.0003	(1.12, 1.48)
Regimen	Insulin and oral Agents	-0.267	0.765	0.1016	-2.632	0.0085	(0.63, 0.93)
	Oral Agents	-0.231	0.794	0.0857	-2.692	0.0071	(0.67, 0.94)
	Regimen	Chisq=11.6208		Df=2	global p-Value=0.003		
Frailty parameter, Theta		0.034		SE (H): 0.0316		p = 0.1394	
penalized marginal log-likelihood = -4804.45 Convergence criteria: parameters = 2.98e-07 likelihood = 1.53e-06 gradient = 3.28e-11 LCV = the approximate likelihood cross-validation criterion in the semi-parametrical case= 3.79188							
Coef: estimated coefficient for each covariates; exp(coef): exponentiation value of coefficient, and se coef(HIH): estimator of standard error ("sandwich estimator") respectively:							

4.2.3 Shared Log-Normal Frailty Model

Similarly, we conducted univariable analysis for the shared log-normal Frailty model. The result suggested that sex, Regimen group, specific drug used and weight were statistically significant at 25% level of significance as presented in the Appendix (Table A4). Sex and Regimen group were the only significant covariates selected from the saturated multiple shared log-normal Frailty model.

Parameter estimates of the Shared Log-Normal Frailty model are presented in Table 6. After controlling for others prognostic factors and accounting for Frailty, the hazard rate of for sex is $\widehat{HR}$=exp ($\hat{\beta}$) =exp (0.254) =1.290. The result revealed that male diabetic patients recover to normal blood sugar level at a rate of about 1.290 times more than female patients. In addition, the Global Chi-Square test suggested that for Regimen group at least one of the Regimen group (Insulin & Oral Agents, and Oral Agent only) is statistically significantly different from the Insulin Agents Only (Reference). The estimated hazard rate for both Regimen group implied that patients who are grouped in the Insulin and oral agents were recovering at a rate of about 0.759 times less than patients who were grouped to Insulin Agents Only, holding others covariates constant and accounting for Frailty. This mean that patients who are grouped to Insulin and Oral agents took longer time to recover to normal blood sugar level as compared the patients grouped to Insulin Agents Only. Similarly, patients who are grouped in the Oral Agents Only were recovering at a rate of about 0.791 times less than patients who were grouped in Insulin Agents Only, holding others covariates constant and accounting for Frailty. This mean that patients who are grouped to Oral Agents Only took longer time to recover to normal blood sugar level as compared the patients grouped to Insulin Agents Only.

Test hypothesis for the variance term of both shared gamma and log-normal Frailty term is given by:

$$H_o: \theta = 0 \quad or \quad H_o: \delta^2 = 0 \qquad vs \qquad H_1: \theta > 0 \qquad or \qquad H_1: \delta^2 > 0$$

The variance of the Frailty term **Theta** and **Sigma Square** are significantly different from zero, meaning that there is heterogeneity between subjects. We can deduce this by using a modified Wald test: W_m (θ) = 0.034/0.0316=1.0759, and W_m (δ^2) = 0.088/0.0157=5.605, with the critical value for a normal one-sided test. The modified Wald test (Wm) is a significance test for the variance of the random effects distribution occurring on the boundary of the parameter space. The usual squared Wald statistic is simplified to a mixture of two distributions and hence the critical values must be derived from this mixture (Molenberghs and Verbeke (2007)). In the case our result we have a p-value that is less than 5% for shared log-normal Frailty but not for shared gamma Frailty distribution. This mean that there is a significant Frailty effect that is, that within subject correlation cannot be ignored for shared log-normal Frailty but not for shared gamma Frailty.

Table 6: Parameter Estimates of Shared Log-Normal Frailty Model.

Variable Names	Category	Coef	exp(coef)	SE coef (HIH)	z	P_value	95%CI for Hazard
Sex	Male	0.254	1.290	0.0762	3.34	0.0008	(1.11, 1.50)
Regimen	Insulin and oral Agents	-0.275	0.759	0.1087	-2.53	0.0113	(0.61, 0.94)
	Oral Agents	-0.235	0.791	0.0914	-2.57	0.010	(0.66, 0.95)
	Regimen	Chisq=10.647		Df=2	global p-Value=0.0049		
Frailty parameter, Sigma Square		0.088		SE (H): 0.0157		p = 9.7975e-09	
penalized marginal log-likelihood = -4800.51 Convergence criteria: parameters = 6.55e-08 likelihood = 2.51e-05 gradient = 1.72e-11 LCV = the approximate likelihood cross-validation criterion in the semi parametrical case=3.79047							
Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, and se coef (HIH): estimator of standard error ("sandwich estimator") respectively.							

4.3. Assessment of model Adequacy

Cox - Snell residuals plot

A plot of the Cox-Snell residuals against the cumulative hazard is presented in Figures 2, 3 and 4. If the model fits the data, the plot of cumulative hazard function against Cox-Snell residuals should be approximately a straight line with slope one. As Hosmer (1998) stated, if the model is the correct one, the points should follow a 45 degree line at the origin. Therefore, the hazard function follows approximately the 45 degree line for the shared log-normal Frailty model than the Cox PH and shared gamma Frailty models. Thus, the Cox-Snell residuals supported that shared log-normal Frailty model fit the data better.

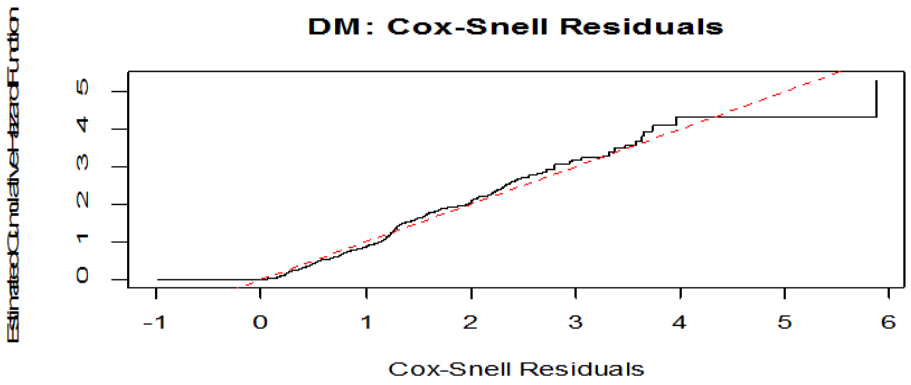

Figure 2: Cox-Snell residuals obtained from Cox-PH model to the DM Data.

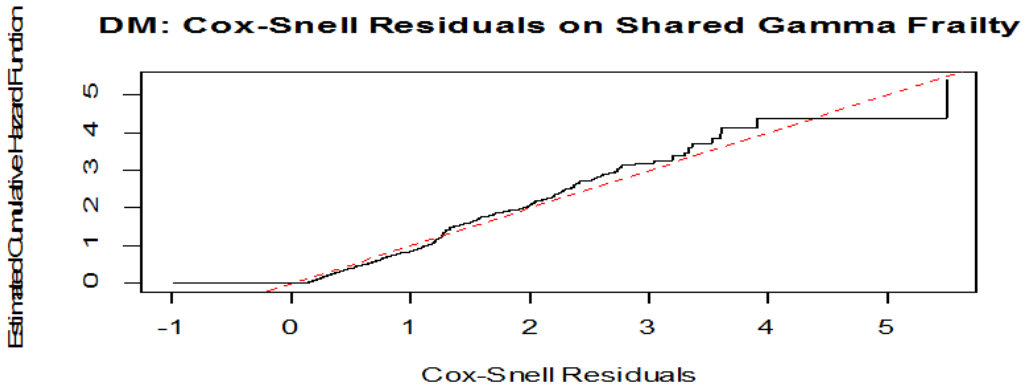

Figure 3: Cox-Snell residuals obtained from shared Gamma Frailty model to the DM Data.

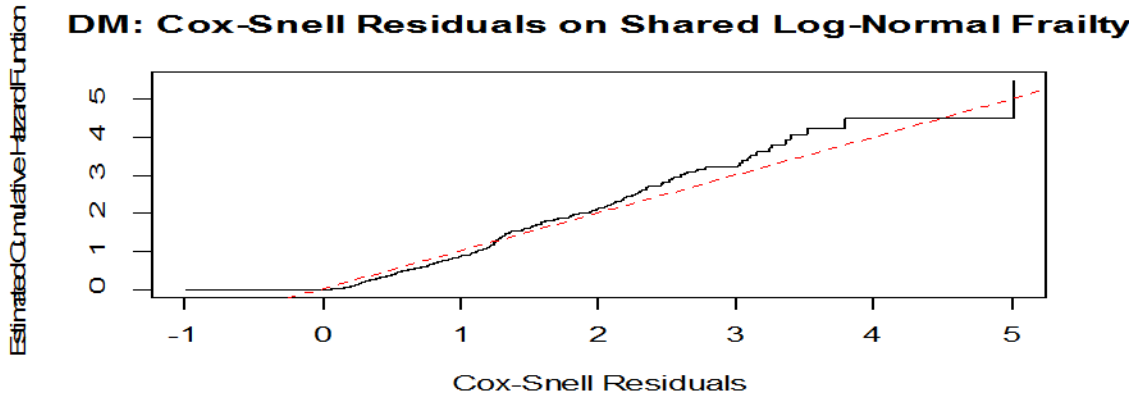

Figure 4: Cox-Snell residuals obtained from shared Log-normal Frailty model to the DM Data.

4.4. Comparison of the Cox-PH and Shared Frailty Models

Efficiency of the fitted models was compared using Penalized marginal log-likelihood and LCV (likelihood cross-validation) criterion. The likelihood cross-validation criterion assesses the goodness of fit of a statistical model (Gray, 1992). In case of parametric approach, LCV is approximately equivalent to Akaike Information criterion (AIC). Lower values of LCV indicate a better fitting model. Table 7 depicts the LCV results of Cox PH, shared gamma Frailty and shared

log-normal Frailty models. The shared log-normal Frailty model was chosen as the best fit of our recurrent events data based on the residual analysis and minimum value of LCV. Although the difference in LCV value of the fitted models was negligible, the Cox-Snell residual plot suggested that the shared log-normal Frailty model fits the data better. The Wald test results indicated that the heterogeneity parameter was significant, implying that there is a significant Frailty effect, or that within subject correlation cannot be ignored.

Table 7:- Comparison of Cox PH and Shared Gamma & Log-normal Frailty Models for DM.

Model	Frailty Parameter	Stander error	penalized marginal log-likelihood	LCV	Rank
Cox-PH	-	-	-4819.2	**3.80041**	3
Shared Gamma Frailty	0.034	0.0316	-4804.45	**3.79188**	2
Shared Log-Normal Frailty	0.088	0.0157	-4800.51	**3.79047**	1

4.5. Discussion

From 347 patients, the majority; 273(78.7%) were Type-II diabetic, 199(57.3%) were females and 199(57.3 %) were grouped to Insulin regimen. The mean weight, age, SBP and DBP at baseline for diabetic mellitus patients were 49 years, 64.92 kg, 127.28 mmHg and 77.83 mmHg respectively.

From the total of 1273 case of events, 923 (72.56%) experienced the event and the rest 350 (27.44%) loss to follow-up from the study. The minimum and maximum recovery times of diabetic were 1 and 318 weeks respectively. The median recovering time of the diabetic patients was 32 weeks in the recurrent events of time to recovery to normal FBS level.

In Cox-PH, shared gamma Frailty and shared log-normal Frailty models of recurrent events of time to recovery of diabetic patients; sex and Regimen group of patients at baseline showed statistically significant association with recurrent events of time to recovery to normal blood sugar level. In univariable and multivariable analysis of standard Cox PH and Frailty models, the sex of patients showed a significant relation indicating better recovery time for male patients than female patients. Our study result showed that female patients were affected by diabetic mellitus for prolonged recovery time as compared to male patients. These findings are consistent with a study conducted that focusing time to first recovery by Abiyot N. (2014).

In this study, sex and Regimen group for three models were found to be statistically significant at 5% level of significance. LCV values of for the standard Cox PH, shared gamma Frailty and for shared log-normal Frailty models were 3.80041, 3.79188 and 3.79047 respectively. According to the results, the shared log-normal Frailty model has the minimum LCV value as compared to that the standard Cox PH and shared gamma Frailty Models, respectively.

The patients who group in the insulin and Oral agents are recovering at rate 0.759 (HR=0.759, 43.15% chance of the Insulin and oral agents group patients recovered as compared to Insulin agents group patients) times of those patients who were grouped in the Insulin agents only, after accounting heterogeneity and holds other covariates constant. Similarly, patients who were grouped in Oral Agents only recovery at a rate 0.791(HR=0.791, 44.17% chance of the patients who are groped in Oral Agents Only recovered as compared to patients who are groped in Insulin Agents Only) times of those patients who were grouped in Insulin Agents Only, after accounting Frailty and holds other factors constant. This means that patients who are grouped to Insulin & oral agents and Oral Agents Only have a prolonged recovery time as compared the patients grouped to Insulin Agents Only.

In Cox-PH, shared gamma Frailty and shared log-normal Frailty models gender was a significant prognostic factor in univariable and multivariable analysis, indicating that the recovery time to normal blood sugar level for males was better than that of female diabetic patients. This means that males affected by diabetic mellitus take shorter time to recover to normal blood glucose level than females. When the Frailty term was ignored, the estimate for β and its standard error was smaller compared to the shared gamma and log-normal Frailty models. This is expected as the Frailty model accounts for the extra variance associated with unmeasured risk factors.

The heterogeneity parameter of θ and $\sigma^2$are estimated to be: 0.034, (SE=0.0316, and P-value=0.1394) and: 0.088, (SE=0.0157 and p=9.7975e-09) for the shared gamma Frailty and shared log-normal Frailty models respectively. The above result suggest not to reject for the gamma Frailty but to reject the log-normal Frailty. This means that, in the shared log-normal Frailty the correlation within cluster patient cannot be ignored. In this study to compare the efficiency of the models the LCV (Likelihood Cross-Validation Criterion), a criterion that assess

goodness of fit a statistical model was used. Shared log-normal Frailty model had relatively the smallest LCV as compared to standard Cox PH and shared gamma Frailty model. Hence, the shared log-normal Frailty model was chosen to analyze recurrent events of time to recovery of diabetic patients when compared to classical Cox PH and shared gamma Frailty in diabetic mellitus dataset.

5. Conclusion and Recommendation

5.1. Conclusion

Diabetes is a group of diseases marked by high or low level of glucose resulting from defects in insulin production, insulin action or both. It can lead to serious complication and premature death but steps to control the disease and lower the risk of complications does exist. Insulin replacement is required for survival. The intensive-therapy regimen was designed to achieve blood glucose values as close to the normal range as possible. In the literature, there are many studies on the field of diabetic, but researchers tend to examine the effects of covariates on patients using logistic regression model and standard survival analysis methods. A model that is becoming increasingly popular for modeling association between individual survival times within subgroups is the use of a Frailty model. In our study, the Frailty represents the total effect on survival of the covariates not measured when collecting information on group of subjects. Most often applied Frailty distribution are the gamma and log-normal distribution.

The main goal of this study was to fit recurrent event models to time-to-recovery for diabetic patients in Menellik II Referral Hospital. The standard Cox PH, shared gamma Frailty and shared log-normal Frailty models were fitted.

The data consisted of 1273 observations for a total of 347 patients. About 72.56% of them experienced the event of interest while the remaining 27.44% of observations did not experience the event of interest throughout the study period.

From the Standard Cox proportional hazard and Frailty models sex and regimen group were found to be statistically significant factors for time to recovery of diabetic patients to normal level of FBS in Menellik II Referral Hospital. The value of the LCV was used to identify the

parsimonious model. Accordingly, the shared log-normal Frailty models provide suitable (appropriate) choice for the life time model of recurrent event time to recovery of the Diabetes Mellitus (DM) as compared to standard Cox proportional hazard without Frailty and Shared gamma Frailty models.

5.2. Limitations of the Study

- The study was conducted based on secondary data which might have incomplete and biased information.
- The study was restricted to adults in the age group >15 and result might not be applicable to infants and children.
- Lack of information on physical activity and diet style that might have contribution on the time to recovery of the DM patients.

5.3. Recommendation

Based on the study finding the following recommendations are forwarded:

- Our result showed sex of patient and regimen group were statistically associated with time to recovery to normal blood sugar level. This calls for actions on improvement of patients' health status based on gender and regimen group.

- Further studies considering other methods of recurrent events time-to-recovery such as marginal, conditional for the calendar time scale (i.e. time to event models) and also gap time scale (time between event models) are recommended.

Reference

1. Abraham Tamirat, Lakew Abebe, Getachew Kirose. (2014): Prediction of Physical Activity among Type-2 Diabetes Patients Attending Jimma University Specialized Hospital, Southwest Ethiopia: Application of Health Belief Model. *Science Journal of Public Health.* Vol. 2, No. 6, pp. 524-531. Doi: 10.11648/j.sjph.20140206.15.
2. Abiyot Negash, Dr. Yehenew Getachew, Mr. Belay Birlie (2014): Modeling Time-to-Frist Recovery of Adult Diabetic Patients: A Case Study of Jimma University Specialized Hospital, Ethiopia.
3. American Diabetes Association (2010): Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care. Jan.* S62–S69, doi: 10.2337/dc10-S062, PMCID: PMC2797383
4. Asrat Agalu Abejew, Abebe Zeleke Belay, and Mirkuzie Woldie Kerie (2014): Diabetic Complications among Adult Diabetic Patients of Dessie Tertiary Hospital in Northeast Ethiopia. *Hindawi Publishing Corporation Advances in Public Health Volume 2015*, Article ID 290920, 7 pages http://dx.doi.org/10.1155/2015/290920.
5. Bahta Y, Fikreyesus Y. (1998): Analysis of admissions to the medical intensive care unit of a teaching hospital in Addis Ababa. *Ethiop Med J,* 26:61-7
6. Blahos J, Kubastova B. (1963): The survey of 11,170 patients treated in the Ras Mekonnen Hospital in Harrar. *Ethiop Med J*, 1:190.
7. Bloom D, Canning D, Sevilla J.: The effect of health on economic growth: A production function approach. World Development. 32(1), 1–13. 20,

8. Collett, D. (2003): *Modeling survival data in medical research (Second edition).* Chapman and Hall/CRC, London.
9. Cox D. (1972*): Regression models and life tables (with discussions).* Journal of the Royal Statistical Society. 34: 187-220.
10. Cox D. R. and Oakes D. (1984): *Analysis of Survival Data". Chapman and Hall, London*.
11. Cox D. R. and Snell E. J. (1968). *A general definition of residuals with discussion. Journal of the Royal Statistical Society*. Series B 30 (1968), 248-275.
12. Connor, H. and Boulton, A. J. M. (1989). Diabetes in Practice. Chichester. Wiley
13. David G. Kleinbaum and Mitchel Klein (2005): *Survival Analysis, Statistics for Biology and Health*. Second Edition, Spring Street, New York, NY, 10013, USA, http://www.sph.emory.edu/ ~dkleinb/surv2.htm.
14. Diabetes (2010): the hidden pandemic and its impact on sub-Saharan Africa. Diabetes Leadership forum Africa.
15. Duchateau L. and Janssen P. (2008): *The Frailty Model.* Springer-Verlag, New York.
16. Dunstan DW, Zimmet PZ, Welborn TA, De Courten MP, Cameron AJ, Sicree RA *et al*.(2002): The rising prevalence of diabetes and impaired glucose tolerance*:* the Australian Diabetes, Obesity and Lifestyle Study. *Diabetes Care;* 25(5):829-834.
17. Elbagir MN, Eltom MA, Elmahadi EM, Kadam IM, Berne C. (1996): A population-based study of the prevalence of diabetes and impaired glucose tolerance in adults in northern Sudan. *Diabetes Care*, 19(10):1126-1128.
18. Emeharole P.O. (2008): Communicable and Non-Communicable Diseases, Control Technology, PP. 162-178.

19. Giday A, Tadesse B. (2011): Prevalence and determinants of hypertension in rural and urban areas of southern Ethiopia. *Ethiop Med J*, 49:139-47.
20. Gill G., English P., Price C., Dedicoat M., and Tesfaye S. (2010): The variable African diabetic phenotype: tales from the north and the south. *12 African Journal of Diabetes Medicine*. Vol. 18 No 2.
21. Grabber A., Elasy T., Quinn D., and Wolff, K. (2002): Improving glycemic control in adults with diabetes mellitus. *South Med Journal*, vol. 95, 684-690.
22. Gray RJ, (1992): Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis, JASA, 87 942-951.
23. Habte-Gabr E., Girma B., Mehrete M., Mehari A., Tekle E., Belachew K. et al.(1976): Analysis of admissions to Gondar Hospital in North-Western Ethiopia, 1971-1972. *Ethiop Med J,* 14:49-59.
24. Hamman R.F., Wing R.R., Edelstein S.L., Lachin M. J., George A. Bray, Linda Delahanty, Mary Hoskin, Andrea M. Kriska, Elizabeth J. Mayer-Davis, Xavier Pi-Sunyer, Judith Regensteiner, Beth Venditti, and Judith Wylie-Rosett (2006): Effect of weight loss with lifestyle intervention on risk of diabetes. *Diabetes Care,* 29:2102–2107.
25. Hosmer DW, Lemeshow S. and May S. (2008): *Applied survival analysis regression modeling of time to event data*. John Wiley and Sons, Inc., New York.
26. Hougaard P. (2000): *Analysis of Multivariate Survival Data*. Springer-Verlag, Newyork.
27. International Diabetes Federation, Third Edition. (2007): The Diabetes Atlas.
28. International Diabetes Federation, Fourth Edition. (2009): The Diabetes Atlas.

29. Internationals Diabetes Federation Fifth Edition. (2011): www.idf.org/diabetes atlas.

30. International Insulin Foundation RAPIA Initiative: http://www.idf.org/webdata/docs/IIFRAPIA_plan.pdf.

31. Issaka Sagara, Roch Giorgi, Ogobara K Doumbo, Renaud Piarroux and Jean Gaudart. Sagara et al. (2014): Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent episodes data. *Malaria Journal*, 13:293: http://www.malariajournal.com/content/13/1/293

32. Joly P., Leteneur D. L., Alioum A., Commenges D., (1999): a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data, Computer Method and Program.

33. King H., Aubert R.E. and. Herman W.H., (1998): Global burden of diabetes, 1995–2025, prevalence, numerical estimates, and projection. *Diabetes Care*, vol. 21:1414–1431.

34. Lainovic D. (1974): Morbidity study in Dire Dawa. *Ethiop Med J,* 12:13-24.

35. Leong, W.Y. (2007): Survival analysis of Diabetic patients in queen Elizabeth hospital, kota kinabalu.

36. Lester FT. (1982): Medical Diseases in the elderly Ethiopians. *Ethiop Med J,* 20:55-61.

37. Liang.K.Y. Self, S.G., Bandeen-Roche, and Zeger, S. (1995): *Some recent developments for regression analysis of multivariate failure time data*. Lifetime Data Analysis, 1:403-416.

38. Mbanya JC, Ngogang J, Salah JN, Minkoulou E, Balkau B.(1997): Prevalence of NIDDM and impaired glucose tolerance in a rural and an urban population in Cameroon. *Diabetologia,* 40(7):824-829

39. Mbanya JC. (2009): Making a difference to global diabetes. *Diabetes Voice,* 54 (3).

40. Melaku Z, Alemayehu M, Oli K, Tizazu G. (2006): Pattern of admissions to the medical intensive care unit of Addis Ababa University Teaching Hospital. *Ethiop Med J*, 44:33-42

41. Misganaw A, Mariam DH, Araya T. *(2012):* The double mortality burden among adults in Addis Ababa, Ethiopia, 2006-2009. *Prev Chronic Dis*, 9: E84.

42. Murray CJ, Lopez AD. (1996): The global burden of disease. World Health Organization. Global burden of disease and injury.

43. Olive D., Buhule, Jonathan O., Odwee and Leonard K. Atuhaire (2007): Determinants of Recovery Time of Diabetic Patients in Uganda. Institute of Statistics and Applied Economics, Makerere University.

44. Pavlica D. (1970): Analysis of medical admissions to the armed forces hospital in Addis Ababa from January 1966 to January 1970. *Ethiop Med J*, 193-200.

45. Pickels, A., Crouchley, R., Simonoff, E.L., Meyer, J., Rutter, M., Hewitt, J. and Silbery, J. (1994): Survival models for developmental genetic data: Age at onset of puberty and antisocial behavior in twins. *Genetic Epidemiology*, 11: 155-170.

46. Self S.G., Liang K., (1987): Asymptotic properties of maximum likelihood estimation and likelihood ratio test under nonstandard condition, JASA, 82 605-610.

47. Rondeau V., (2010): Statistical models for recurrent events and death: Application to cancer events. Mathematical and Computer Modelling 52 (2010) 949-955. France

University Victor Segalen Bordeaux2, INSERM, CR897 (Biostatistics). doi:10.1016/j.mcm.2010.02.002.

48. Ramsay J.O., (1988): Monotone Regression splines in action, Statistical Science, 3, 425-461.
49. T/Kiros Giday, Huruy Aseffa and Abadi Kidanemariam (2014), studied on the assessment of risk factors associated with Type II Diabetes Mellitus in Central Zone of Tigray, North Ethiopia.
50. Teklu B. (1982): Chronic disease prevalence in Ethiopian bank employees. *Ethiop Med J,* 20:49-54.
51. Temesgen Fiseha, Mehidi Kassim, Tilahun Yemane. (2014): Prevalence of Chronic Kidney Disease and Associated Risk Factors among Diabetic Patients in Southern Ethiopia. *American Journal of Health Research*. Vol. 2, No. 4, pp. 216-221. Doi: 10.11648/j.ajhr.20140204.28.
52. Ullah S, Gabbett TJ, Finch CF. (2014): Statistical modelling for recurrent events: an application to sports injuries. *Journal Sports Med:* 48: 1287 – 1293
53. Virginie Rondeau, Daniel Commenges, and Pierre Joly: Maximum penalized likelihood estimation in a gamma-frailty model. Lifetime Data Analysis Additional article information.
54. Vivian C. Tuei1, Geoffrey K. Maiyoh3Chung-Eun Hal (2010): Type II diabetes mellitus and obesity in sub-Saharan Africa, Diabetes/Metabolism Research, and Reviews Review Article. *Diabetes Metab Res Rev; 26: 433– 445. Published online19 July 2010 in Wiley Online* Library: www.wileyonlinelibrary.com.

55. World Health Organization (2001): The World Health Report 2001. Geneva, World Health Organization.

56. World Health Organization (2016). Diabetes estimates and projections. Diabetes fact sheet reviewed June 2016. http://www.who.int/mediacentre/factsheets/fs312/en/ accessed on July 2016**.**

57. Yashin, A.I., Vaupel, J.W., and Jachine, I.A. (1995): *Correlated individual Frailty: An advantageous approach to survival analysis of bivariate data.* Mathematical population studies, 5:145-149.

58. Yemane T, Belachew T, Asaminew B, Befekadu O. (2007): Type II diabetes mellitus in Jimma Town, southwest Ethiopia. *Ethiop J Health Sci*: 17.

Appendix

Table A1: covariate are described with their values (codes) as follows

No	Variables	Description	Value? Codes
1	sex	Sex of patients	0= Female, 1=Male
2	Type.DM	Diabetic Types	0= Type I DM, 1= Type II DM
3	PastMed.H	Past Medical History	0=Yes, 1=No
4	FamilyH	Family History	0=Yes, 1=No
5	Complication	Is there any Complication Status	0=Yes, 1=No
6	Martst	Marital status	0=Married, 1=Single
8	Education_Satstus	Education Status of patients	0=educated, 1=uneducated
9	Employee_Status	Employee Status of patients	0=employee, 1=unemployed
10	Regimen	Regimen type	0=Oral Agents, 1=Insulin Agents, 2=Oral and Insulin Agents
11	Spdrty	Specific drug(treatment) type order by Physician	0=Doanied, 1=HCT, 2=Metformin, 3= Monotend, 4=Lute, 5= Regular, 6=all oral
12	SBP	Systolic Blood Pressure in Mm/Hg	0=<110(below), 1=110-130(normal), 2=>130(High)
13	DBP	Diastolic Blood Pressure in Mm/Hg	0=<60(below), 1=60-80(normal), 2=>80(High)
14	Agec	Age of Patient in category	0=15-29, 1=30-44, 2=45-59, 3=60-74, 4= above 74
15	Wight	Weight of patients in Kg	Continuous Variable

Table A2: Parameter Estimates of Cox PH Model for Univariable Analysis

No.	Covariates	Category	coef	Exp(coef)	SE coef (H)	SE coef(HIH)	p
1	Sex	Male	0.283	1.327	0.0663	0.0663	0.00002
2	Agec	>74	0.074	1.077	0.2364	0.2364	0.754
		15-29	0.110	1.117	0.1025	0.1025	0.282
		30-44	-0.049	0.952	0.0915	0.0915	0.593
		60-74	0.082	1.086	0.0842	0.0842	0.328
		agec		Chisq=3.094, Df=4		Global P_value=0.542	
3	Type of DM	Type II DM	-0.076	0.926	0.0789	0.0789	0.335
4	Past Medical History	Yes	-0.004	0.996	0.0728	0.0728	0.958
5	Family History	Yes	0.076	1.079	0.0828	0.0828	0.359
6	Complication	Yes	-0.025	0.975	0.0702	0.0702	0.704
7	Marital Status	Single	0.012	1.012	0.0679	0.0679	0.859
8	Educational Status	uneducated	-0.026	0.974	0.0771	0.0771	0.732
9	Employee Status	unemployed	-0.022	0.978	0.0659	0.0659	0.7398
10	Regimen	Insulin and Oral Agent	-0.328	0.721	0.0954	0.0954	0.0006
		Oral Agents	-0.225	0.798	0.0818	0.0818	0.0058
		Regimen		Chisq=16.039, Df=2		Global P_value=0.00033	
11	Spdrty	Doanied	-0.201	0.818	0.0813	0.0813	0.0134
		HCT	0.131	1.139	0.4105	0.4105	0.7501
		Metformin	-0.048	0.953	0.3118	.03118	0.878
		Lute	0.084	1.088	0.3058	0.3058	0.783
		Regular	0.388	1.474	0.2549	0.2549	0.128
		All oral	-0.352	0.704	0.1015	0.1015	0.0005
		Spdrty		Chisq=19.3327, Df=6		Global P_value=0.0036	
12	SBP	Below(,=110)	0.019	1.019	0.0968	0.0968	0.841
		High(>=130)	0.013	1.013	0.0753	0.0753	0.858
		SBP(news)		Chisq=0.0489 , Df=2		Global P_value=0.976	
13	DBP	Below(<=60)	-0.013	0.987	0.2640	0.2640	0.959
		High(>=80)	0.005	1.005	0.0686	0.0686	0.946
		DBP		Chisq=0.0086, Df=2		Global P_value =0.996	
14	Wight		0.003	1.003	0.0029	0.0029	0.249

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direc t and the Hessian ("sandwich estimator") respectively.

Table A3: Parameter Estimates of Shared Gamma Frailty Model for Univariable Analysis

No.	Covariates	Category	coef	Exp(coef)	SE coef(HIH)	P_value
1	Sex	Male	0.285	1.329	0.0709	0.00006
	Frailty parameter, Theta=0.047		SE (H) = 0.0329		P_value=0.0782	
2	Agec	>74	0.062	1.064	0.2613	0.812
		15-29	0.111	1.118	0.1130	0.323
		30-44	-0.048	0.953	0.0999	0.633
		60-74	0.088	1.092	0.0925	0.339
		agec	Chisq=2.6741, Df=4		Global P_value=0.614	
	Frailty parameter, Theta=0.067			SE (H) = 0.0344		P_value=0.026
3	Type of DM	Type II DM	-0.076	0.927	0.0871	0.383
	Frailty parameter, Theta=0.069			SE (H) = 0.0346		P_value=0.0229
4	Past Medical History	Yes	-0.009	0.991	0.0803	0.908
	Frailty parameter, Theta=0.071			SE (H) = 0.0346		P_value=0.0212
5	Family History	Yes	0.071	1.074	0.0914	0.436
	Frailty parameter, Theta=0.068			SE (H) = 0.0345		P_value=0.0238
6	Complication	Yes	-0.24	0.975	0.0772	0.750
	Frailty parameter, Theta=0.069			SE (H) = 0.0346		P_value=0.0217
7	Marital Status	Single	0.018	1.018	0.0779	0.812
	Frailty parameter, Theta=0.070			SE (H) = 0.0346		P_value=0.0214
8	Educational Status	uneducated	-0.024	0.976	0.0849	0.773
	Frailty parameter, Theta=0.069			SE (H) = 0.0346		P_value=0.0214
9	Employee Status	unemployed	-0.022	0.978	0.0725	0.7548
	Frailty parameter, Theta=0.069			SE (H) = 0.0346		P_value=0.0215
10	Regimen	Insulin and Oral Agent	-0.333	0.717	0.1019	0.0011
		Oral Agents	-0.232	0.793	0.0873	0.0078
		Regimen	Chisq=14.5952, Df=2		Global P_value=0.00068	
	Frailty parameter, Theta=0.050			SE (H) = 0.0331		P_value=0.0645
11	Spdrty	Doanied	-0.207	0.813	0.0866	0.0170
		HCT	0.178	1.195	0.4409	0.6868
		Metformin	-0.061	0.941	0.3317	0.8550
		Lute	0.128	1.137	0.3277	0.6956
		Regular	0.379	1.460	0.2687	0.1587
		All oral	-0.353	0.703	0.1076	0.001
		Spdrty	Chisq=17.633, Df=6		Global P_value=0.00722	
	Frailty parameter, Theta=0.047			SE (H) = 0.0328		P_value=0.0777
12	SBP	Below(,=110)	0.033	1.033	0.1065	0.758
		High(>=130)	0.024	1.025	0.0829	0.7684
		SBP	Chisq=0.123 , Df=2		Global P_value=0.94	
	Frailty parameter, Theta=0.071			SE (H) = 0.0346		P_value=0.0209
13	DBP	Below(<=60)	-0.006	0.994	0.2875	0.984

		High(>=80)	0.012	1.012	0.0755	0.870
		DBP	Chisq=0.029, Df=2		Global P_value =0.986	
	Frailty parameter, Theta=0.07			SE (H) = 0.0346		P_value=0.0212
14	Wight		0.004	1.004	0.0032	0.0250
	Frailty parameter, Theta=0.064			SE (H) = 0.0339		P_value=0.0303
Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, and se coef (HIH): estimator of standard error ("sandwich estimator") respectively.						

Table A4: Parameter Estimates of Shared Log-Normal Frailty Model for Univariable Analysis

No.	Covariates	Category	coef	Exp(coef)	SE coef(HIH)	P_value
1	Sex	Male	0.283	1.327	0.0749	0.00016
	Sigma Square=0.09		SE (H) = 0.0168		P_value=4.6159e-08	
2	Agec	>74	0.068	1.071	0.2699	0.7997
		15-29	0.111	1.117	0.1166	0.3436
		30-44	-0.046	0.955	1.029	0.6651
		60-74	0.095	1.099	0.0955	0.3212
		agec	Chisq=2.6088, Df=4		Global P_value=0.625	
	Sigma Square=0.093			SE (H) = 0.0187		P=3.6024e-07
3	Type of DM	Type II DM	-0.078	0.925	0.0899	0.385
	Sigma Square=0.093			SE (H) = 0.0191		P= 5.3296e-05
4	Past Medical History	Yes	-0.007	0.993	0.0828	0.936
	Sigma Square=0.094			SE (H) = 0.0192		P= 5.3296e-07
5	Family History	Yes	0.071	1.074	0.0945	0.450
	Sigma Square=0.0932			SE (H) = 0.0191		P= 4.9633e-07
6	Complication	Yes	-0.26	0.974	0.0796	0.7396
	Sigma Square=0.0935			SE (H) = 0.0192		P= 5.694e-07
7	Marital Status	Single	0.018	1.019	0.0770	0.810
	Sigma Square=0.0934			SE (H) = 0.0195		P= 5.719e-07
8	Educational Status	uneducated	-0.023	0.977	0.0874	0.7887
	Sigma Square=0.0935			SE (H) = 0.0192		P= 5.6457e-07
9	Employee Status	unemployed	-0.023	0.977	0.0746	0.7598
	Sigma Square=0.0935			SE (H) = 0.0192		P= 5.6371e-07
10	Regimen	Insulin and Oral Agent	-0.338	0.713	0.1072	0.0016
		Oral Agents	-0.235	0.791	0.0915	0.010
		Regimen	Chisq=13.5861, Df=2		Global P_value=0.0011	
	Sigma Square=0.090			SE (H) = 0.0169		P= 5.7198e-08
11	Spdrty	Doanied	-0.212	0.809	0.0914	0.020
		HCT	0.175	1.191	0.4707	0.7107
		Metformin	-0.049	0.952	0.3474	0.8864
		Lute	0.162	1.175	0.3410	0.6357
		Regular	0.371	1.449	0.2809	0.1872
		All oral	-0.353	0.703	0.1076	0.001
		Spdrty	Chisq=16.0895, Df=6		Global P_value=0.0133	
	Sigma Square=0.089			SE (H) = 0.0165		P= 2.8905e-08
12	SBP	Below(,=110)	0.029	1.029	0.1096	0.7878
		High(>=130)	0.026	1.026	0.0854	0.7603
		SBP	Chisq=0.1124 , Df=2		Global P_value=0.945	
	Sigma Square=0.0935			SE (H) = 0.0193		P= 5.9544e-07
13	DBP	Below(<=60)	-0.006	0.994	0.2954	0.9836
		High(>=80)	0.009	1.009	0.0777	0.9036
		DBP	Chisq=0.0163, Df=2		Global P_value =0.992	
	Sigma Square=0.0935			SE (H) = 0.0192		P= 5.7764e-07
14	Wight		0.003	1.003	0.0033	0.0249
	Sigma Square=0.0935			SE (H) = 0.0192		P= 5.694e-07

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, and se coef (HIH): estimator of standard error ("sandwich estimator") respectively