MODELLING OF SELECTED SOCIO-ECONOMIC AND DEMOGRAPHIC PREDICTORS OF DIABETIC KIDNEY DISEASE AMONG DIABETIC PATIENTS: COMPARATIVE ANALYSIS OF COX REGRESSION AND SUPPORT VECTOR MACHINE MODELS

GRACE MAKENA NJOKA

A Thesis Submitted to the Graduate School in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science in Applied Statistics of Chuka University.

CHUKA UNIVERSITY

OCTOBER, 2024

DECLARATION AND RECOMMENDATION

Declaration

This Thesis is my original work and has not been presented for the award of a degree or diploma in any other university or institution.

Signature.....Date.....24/10/2024.....

Njoka Grace Makena
SM18/58133/22

Recommendation

This Thesis has been examined, passed and submitted with our approval as university supervisors.

Signature.....Date.....24/10/2024.....

Prof. Moses M. Muraya, PhD
Chuka University

Signature.....Date.....24/10/2024.....

Dr. Elizabeth W. Njoroge, PhD
Chuka University

DEDICATION

This research is dedicated to the Njoka's family, who have given me the drive and inspiration to keep going.

ACKNOWLEDGEMENT

First, I would like to express my heartfelt gratitude to God for His unwavering grace, which has made this academic journey possible. I extend my deepest thanks to my esteemed supervisors, Prof. Moses M. Muraya and Dr. Elizabeth W. Njoroge, whose expertise and insights have been invaluable. I am honoured to have had the chance to learn from them and work under their mentorship and supervision. My heartfelt appreciation to my lovely parents, Mr. and Mrs. Armstrong Njoka for their steadfast love, support, encouragement and belief in me, not forgetting my dear siblings Timothy and David, and my daughter Natasha Mwende who have also been the bedrock of my milestones. I owe these persons a debt of gratitude that words may not fully express and I am immensely blessed to have each of them. Lastly, I wish to thank my lecturers in the Department of Physical Sciences at Chuka University for their support and the resources they provided, which have been instrumental in my success.

ABSTRACT

Diabetic kidney disease (DKD) accounts for one in three adults with diabetes and is a significant trigger for mortality among diabetic patients globally. Traditional predictive models of DKD in diabetic patients have mainly been based on patients' clinical health histories but have overlooked socio-economic factors that may also be integral in DKD prevalence. This study aimed to develop an improved predictive model that considers the effects of socio-economic, demographic, and behavioural factors on the survival rates of diabetic patients prior to developing DKD. A retrospective survey design was conducted among 756 diabetic patients at Meru Teaching and Referral Hospital and Kerugoya Level 5 Hospital in Kenya. Patients' records and semi-structured questionnaires were utilised for data collection. The extracted data were entered into Excel and analysed using R software. The Cox regression and Support Vector Machine (SVM) were employed to identify the predictors of DKD and to gauge someone's risk of developing DKD over time based on their socio-economic characteristics. The research data were randomly split into a training set (70%) and a test set (30%) for developing the two models and identifying predictors of DKD. Age at diagnosis, history of cardiovascular disease, alcohol use, financial hardships, employment status, level of education, and gender were identified as significant predictors associated with DKD. The study found that the SVM model had a slightly higher C-index (0.7753) in comparison with the Cox model (0.770), indicating that SVM model was marginally more accurate in predicting DKD than the Cox model. Therefore, prompt policy changes and effective strategies in public health or clinical practice should be designed based on the identified socio-economic predictors and the developed models in an effort to prevent the development of DKD in diabetic patients.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS ANDABBREVIATIONS

AUC	Area Under the Curve
BMI	Body Mass Index (Kg/M^2)
CKD	Chronic Kidney Disease
C-index	Concordance Index
DKD	Diabetic Kidney Disease
eGFR	Estimated Glomerular Filtration Rate
ESKD	End Stage Kidney Disease
LRT	Log-Rank Test
SVM	Support Vector Machine

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Chronic kidney disease (CKD) is one of the long-term complications which can arise from diabetes (Thomas *et al.*, 2015). Approximately 843.6 million people worldwide are affected by CKD (Jager *et al., 2019*; Global Burden of Disease [GBD] Diabetes Collaborators, 2023). This translates to a global burden of 12%. It is a dominant cause of worldwide mortality *(*Rhee & Kovesdy, 2015). According to Abd *et al.,* (2018), prevalence of kidney disease in Africa is 10.1%. About 5 million individuals in Kenya suffer from chronic kidney disease (CKD) (Saya, 2023). This corresponds to a prevalence of 9%.

Previous estimates of the prevalence of CKD in Kenya have been made for a several illnesses, such as rheumatoid arthritis, HIV, type 2 diabetes and heart failure (Kairu, 2015). Diabetic kidney disease (DKD) is a type of kidney condition that arises from long-term complications of diabetes, imposing the highest burden in terms of both the financial cost and on the daily life of the patients (Thomas *et al.*, 2015; Rabkin, 2003). It is the main factor behind chronic kidney disease (CKD) and its progression to end-stage kidney disease in developed countries (Rabkin, 2003). However, little is known about the wide range of risk factors for DKD in developing countries. According to estimates, DKD affects between 30% and 40% of diabetic patients (Unamath & Lewis, 2018).

Survival analysis is a statistical field focused on methods for assessing the probability of an event occurring over time. When the distribution of survival times is unknown, semi-parametric techniques are utilized, while parametric models are suitable when the distribution is established. In health sciences, the Cox proportional hazards regression model is the most widely used semi-parametric model due to its reliance on fewer assumptions in comparison with parametric models (Georgousopoulou *et al.,* 2015). Selingerova *et al.* (2021) note that semi-parametric models incorporate parametric approaches to evaluate the effects of covariates, while non-parametric methods are used for estimating baseline hazards.

The primary benefit of the Cox regression is that does not necessitate a specific statistical distribution for survival time. Its key assumptions is that the effect of variables on the survival outcome remains constant as time passes. This leads to the proportional hazards assumption that states that the hazard rates for two individuals will change in a consistent proportion as time advances. In one dimension, this assumption is easily verifiable; however, in higher dimensions, this verification becomes more challenging. Furthermore, because the proportional hazards model uses partial likelihood for parameter estimation, it becomes more difficult to fit when there are more covariates than individuals (Lee &Wang, 2003; Kleinbaum & Klein, 2012).

Support vector machines (SVMs) offer a different approach compared to the proportional hazards model. Introduced by Vapnik in 1995, SVMs are a machine learning approach designed for binary outcomes. Cervantes *et al.* (2008) highlight their robust theoretical foundation and exceptional classification accuracy, particularly in high-dimensional classification tasks. Unlike the Cox proportional hazards regression, machine learning methods can identify and address higher-order interactions and nonlinear relationships without relying on parametric or semiparametric assumptions (Hu & Steingrimsson, 2017).

Currently new treatments are becoming available to treat kidney disease (Pöhlmann *et al*., 2022). DKD models may be useful in informing clinical and economic decisions regarding these new treatment options. Existing models offer different modelling strategies. However, there is need to improve their accuracy. Moreover, data available may vary depending on underlying characteristics. Model parameters may not always be evident in the existing literature.

Research on predicting the onset of DKD remains limited and has primarily relied on data from clinical trials or diverse groups with various causes of CKD (Dunkler *et al.,* 2015, Jiang *et al.,* 2020). These studies mainly emphasize clinical and genetic factors that influence DKD. Additionally, metabolic conditions are significant risk factors for CKD (Kalantar-Zadeh *et al.,* 2021). Moreover, several genetic factors, including hypertension,

age, gender, dyslipidaemia, hyperglycaemia, cardiovascular disease, and smoking, have been associated with elevated risk of developing DKD (Pacilli et al., 2017).

However, there are still many unanswered questions regarding the pathophysiology of DKD due to its complexity. Numerous studies are increasingly recognizing that socioeconomic factors can contribute to chronic kidney disease (CKD) not just by modifying existing risk factors, but also as independent risk factors in their own right (Weldegiorgis *et al.,* 2020). This indicate that it is necessary to develop model for specific population and thus generalized models may not provide accurate prediction across different populations. In addition, DKD often develops slowly and with few symptoms with majority of cases being detected at an advanced stage. Hence there was need to develop predictive models with higher precision in determining DKD occurrence.

Previous research has investigated the link between socioeconomic factors and medical status in diabetic patients (Walker *et al.,* 2020; Berkowitz *et al.,* 2015). Nonetheless, the studies did not set out to determine how they affected DKD. Studies investigating the connection between financial difficulties and complications arising from diabetes, especially DKD are inadequate. The current study looked at the connection between incident DKD in adults with diabetes and socioeconomic and demographic factors in order to close this knowledge gap.

When there are few or moderately many predictors, current risk prediction models for DKD created using conventional regression models such as logistic or linear regression usually excel. However, where there are many variables, they have a tendency to overfit. (Sabanayagam *et al.,* 2023). In contrast, the Cox regression fails to assume a constant hazard function or any specific distribution. It assumes that regression portion of the model is entirely parametric (Harell, 2015). This model also helps eliminate variables that have minimal or no impact on survival rates, resulting in a refined model that focuses on factors significantly influencing the event of interest. Additionally, survival curves and differences in survival among independent variables are employed making use of Kaplan-Meier curves and log-rank tests (Bradburn *et al.,* 2003).

Both traditional methods and machine learning techniques have been explored to determine the key variables or features to incorporate into a model. (Schroeder *et al*., 2017, Dovgan *et al*., 2020). Some studies have combined regression and machine learning approaches by first identifying significant variables through regression analysis, then incorporating these in developing a prediction model (Schroeder *et al.,* 2017, Dovgan *et al*., 2020, Krishnamurthy *et al*., 2021). The performance of a model is reduced if there are no significant variables or if it includes features that are insignificant (Krishnamurthy *et al*., 2021).

1.2 Statement of the Problem

Diabetic kidney disease (DKD) is increasingly viewed as a major global health issue, significantly affecting the financial and emotional well-being of patients and their families. Various studies have explored the prevalence of DKD and highlighted several predictive factors related to the disease. Nonetheless, many of the current models do not consider the impact of socioeconomic elements like marital status, financial difficulties, and education level, which diminishes their predictive effectiveness. Recognizing the risk factors associated with DKD is crucial in effective treatment and management of the illness. The ability to accurately forecast survival rates of diabetic patients before the development of DKD is vital for improving outcomes and reducing morbidity and mortality in this group. Thus, there is a pressing need to create a statistical model that improves predictive accuracy for survival rates in diabetic patients by incorporating unconventional variables such as marital status, educational background, financial challenges, living situation, and substance use, including alcohol and tobacco.

1.3 Objectives of the Study

1.3.1 Broad Objective

The general objective of this study was to develop an improved predictive model which considers effects of socio-economic and demographic factors on the survival rate of diabetes patients prior to developing diabetic kidney disease using Cox regression and Support Vector Machine.

1.3.2 Specific Objectives

i. To determine the predictors associated with the development of DKD among individuals with diabetes
ii. To develop a predictive model that estimate an individual's risk of developing DKD over time based on their socio-economic characteristics
iii. To compare the performance of DKD prognostic prediction models based on Cox Regression and Support Vector Machine (SVM).

1.4 Research Questions

i. Which are the predictors associated with the development of DKD among individuals with diabetes?
ii. Can a predictive model that estimate an individual's risk of developing DKD over time based on their socio-economic characteristics be developed?
iii. Is there a difference in the performance of DKD prognostic prediction models based on Cox Regression and Support Vector Machine (SVM)?

1.5 Significance of the Study

Over the past thirty years, the incidence of diabetic kidney disease has consistently risen, and this trend is expected to persist worldwide. The majority of people with DKD in its early stages are either asymptomatic or show non-specific symptoms, which can lead to missed diagnoses. The model developed may predict the risk for DKD development upon diabetes diagnosis and further analyse the effects of various socio-economic, behavioural factors and other predominant factors.

The developed model may assist clinicians in healthcare facilities in enhancing decision-making and enabling early risk prediction for diabetic kidney disease (DKD). This proactive approach can facilitate timely interventions, ultimately improving patient outcomes. Early prediction of DKD could lead to therapeutic interventions and lifestyle modifications for diabetic patients, helping to prevent the progression of the illness to more advanced stages, reduce dependency on dialysis, and lower healthcare costs. If

adopted the model will also be used to advice the government and policy makers on need for supporting the diabetes patients through improving the health facilities to have more equipment, financial assistance and employing trained staff to advice and help the patients in order to control prevalence of DKD.

Results from the comparing predictive performance between the developed Cox regression and Support Vector Machine (SVM) models will help widen current knowledge in statistics.

CHAPTER TWO

LITERATURE REVIEW

2.1 Incidence and Prevalence of Diabetes and Diabetic Kidney Disease

2.1.1 Incidence and Prevalence of Diabetes

The World Health Organization (WHO) defines diabetes as a chronic condition caused by either the pancreas making insufficient insulin or the body being unable to effectively use the insulin that is available. Insulin helps in regulating blood sugar levels. According to the American Diabetes Association (2010), diabetes is primarily categorized into two types: type 1 diabetes, defined by absolute insulin deficiency, and type 2 diabetes, which involves insulin resistance and a relative deficiency in insulin production.

Nearly 451 million people worldwide are diabetic and the number is anticipated to increase to 693 million by 2045 (Cho *et al.,* 2018). Furthermore, diabetes is thought to account for over 9.9% of global all-cause mortality, responsible for approximately 5 million deaths. In Africa, diabetes presents a significant healthcare burden, notably in low- and middle-income countries that often lack advanced healthcare infrastructure, where the highest increases in diabetes prevalence are anticipated (Federation, 2019, Zhou *et al.,* 2016). In Kenya the situation regarding diabetes prevalence and management is not clear given the paucity of relevant data. Mwaura (2024) estimates the incidence of diabetes in Kenya to be around 4.5%.

Approximately 33.3% of individuals with diabetes may not be diagnosed, particularly in nations which are developing. They have limited resources, thus the disease is frequently not discovered until complications arise (Wens *et al.,* 2005). In many populations around the world, the disease is now starting much earlier (Uusitupa, 2002).

2.1.2 Incidence and Prevalence of Diabetic Kidney Disease

A review by Jager *et al*. (2019) indicates that the prevalence of chronic kidney disease CKD globally is 11.1%, which translates to nearly 843.6 million people affected by this condition worldwide (11.1% × world population, or prevalence × 11.1% × 7.6 billion). As a result, one in every eleven persons has CKD to some extent. It is a significant global

health issue (Mills *et al.,* 2015). However, the prevalence of CKD varies annually, across studies, across cultures, and within countries (Bikbov *et al.,* 2020).

Diabetes frequently results in diabetic kidney disease (DKD), a chronic disease. It is a significant but little-acknowledged factor in the worldwide disease burden. Notably, 50% of cases of end-stage kidney disease (ESKD) and CKD worldwide are caused by diabetic kidney disease (Tuttle, 2014). According to projections, over 70% of patients will reside in developing nations by 2030. (Yirsaw, 2012).

The prevalence of diabetic kidney disease (DKD) among diabetic persons varies significantly across different countries, with rates ranging from 27% in China to 84% in Tanzania (Afkarian *et al.,* 2013; Parving *et al.,* 2006). The number of people affected by DKD is projected to increase alongside the rising prevalence of diabetes. A systematic review on global DKD and its predictors highlights that a history of type 2 diabetes mellitus a key predictive factor for developing the condition (Fenta *et al*., 2023).

2.2 Predictors of Diabetic Kidney Disease

Previous studies have identified several clinically relevant factors that may expedite the progression of DKD, including age, urine albumin-creatinine ratio, haemoglobin A1c levels, high-density lipoprotein, and insulin resistance (Nakashima *et al.,* 2021, Liu *et al.,* 2021). Additional significant clinical factors include body mass index (BMI), triglyceride levels, smoking, hypertension and the use of specific antidiabetic and antihypertensive medications (Mu *et al.,* 2023). However, this study focuses on the demographic, socioeconomic, and behavioural factors that may also affect the progression of DKD.

Diabetic kidney disease prevalence increases significantly with increase in age, in both men and women (Hoogeveen, 2022). Various studies have shown consistent results that age is a major factor (Lin *et al.,* 2014 and Geletu *et al.,* 2018). Consequently, screening for DKD in the elderly population is a crucial strategy for implementing timely interventions.

Research examining the relationship between gender and the onset of DKD in individuals with diabetes has been insufficiently conducted, yielding conflicting results, and it is still unclear whether gender differences exist. According to Shen *et al.* (2017), women are more prone to develop diabetic end-stage renal disease than men. On the other hand, men with pre-diabetes and newly diagnosed diabetes have an enhanced risk of developing CKD, according to Parizadeh *et al*. (2019).

There is a close relationship between kidney disease and cardio-vascular diseases (CVD). When one organ is sick, the other becomes dysfunctional, which eventually results in the failure of both organs (Liu, 2014). Consequently, a diabetic patient's risk of developing DKD may increase if they have CVD. Albuminuria is thought to be a predictive indicator of either renal or cardiovascular risk, or both (Mule *et al.,* 2017).

Increased renal plasma flow, glomerular hyper filtration, and abnormal albuminuria are linked to severe obesity but these conditions can improve with weight loss. (Chagnac *et al*., 2003). Tapp *et al.* (2004), indicate a link between a high Body Mass Index (BMI) and elevated hazards of DKD. Furthermore, diet and weight loss in diabetic patients may improve kidney function and lower proteinuria (Saiki *et al.,* 2005).

Patients with hypertension have a six-fold increased risk of contracting diabetic nephropathy in comparison with individuals without the condition (Tekalign *et al.,* 2023). According to Verma *et al.* (2016), hypertension is a separate risk factor for DKD. Patho-physiologically, hypertension and DKD are interrelated; chronic hypertension can worsen kidney function, while declining kidney health can negatively affect blood pressure control (Buffet & Ricchetti, 2012).

There is limited research connecting marital status to CKD. One study found that patients with kidney disease who are in unhappy marriages or who face significant marital conflict could suffer serious health issues, including increased mortality (Cohen *et al.,* 2007). Among middle-aged and older individuals, single people are associated with a greater incidence of CKD compared to those married (Chen *et al.,* 2020).

Social economic factors and specifically financial hardships have been associated in increased prevalence of DKD. Funakoshi *et al*. (2017) revealed that individuals with type 2 diabetes from middle- or low-income backgrounds have a higher likelihood of developing nephropathy. Low socioeconomic status is linked to a heightened hazard of progressing to end-stage kidney disease, which necessitates dialysis or a kidney transplant.as demonstrated by Ke *et al.* (2019). Individuals with limited socioeconomic status and education levels are more significantly affected by DKD. (Duru *et al.,* 2018). Financial deprivation must be addressed as a risk factor for DKD because this is a global issue.

2.2.1 Effect Socio-Economic Factors

While some studies have not shown an association between socio-economic factors and contraction of DKD among diabetic patients (Willers *et al.,* 2018; Bihan *et al.,* 2005), others indicate that these socio-economic factors are at least associated with microalbuminuria, which is an initial indicator of DKD. (Wolf *et al.,* 2011; Bihan *et al.,* 2012). Therefore socio- economic factors may have an impact on diabetic patients that may lead to DKD. There are varied mechanisms by which socio economic factors lead to diabetes related kidney disease.

Brown *et al*. (2004) define "access to healthcare" as encompassing both the presence of healthcare services and the extent to which those services are utilized. Even though healthcare may be available to everyone, access may be restricted by things like lengthy waiting lists, the availability of public transportation to services, the need for specialized care, and most importantly - diabetic technology. According to a recent study, there are notable variations in DKD management across various socioeconomic groups (Phillips *et al.,* 2024).

Higher educated individuals might be better able to comprehend their condition, practice self-care, and ultimately achieve good metabolic control (Saeed *et al.,* 2022). This could lead to more complications from diabetes. Diabetes self-care behaviours are negatively

impacted by the fact that individuals with diabetes who are less educated frequently report being less satisfied with their care (Gajewska *et al.,* 2020). This could lead to DKD progression.

A study by Bihan *et al.* (2012) observed that diabetic individuals who come from low-income backgrounds score lower on quality of life measures. This is understandable, as it may create additional financial pressure, which can be worsened by the challenges of managing diabetes, coping with complications, and the potential impact on one's ability to work or earn an income. This could lead to more complications from diabetes, such as DKD.

2.3 Survival Analysis

Survival analysis concentrates on modelling the duration until a specific event occurs, for instance disease recurrence or fatality. "Survival time" refers to the period from an initial event, like a diagnosis or treatment, until the occurrence of the relevant event the research is interested in. The term "failure" is often used to describe this event.

Kleinbaum and Klein (2012) define survival analysis as all the statistical techniques that prioritize survival time as the main variable. This involves examining the time interval from a defined starting point, such as a diabetes diagnosis, to a final event, such as the onset of CKD (Bradburn *et al.,* 2003).

Survival analysis is based on the Survival Function S(t), which represents the probability that the true survival time (T) surpasses a defined time (t). According to Xie & Liu (2005), this function is calculated as follows:

$$S(t) = P[T > t] \quad \ldots\ldots (1)$$

If T is considered a continuous random variable with a cumulative distribution function F(t)) = Pr {T≤ t} and a probability density function f(t), the probability that the event has occurred by time is:

$$f(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < +\Delta t)}{\Delta t} \quad \ldots\ldots (2)$$

The failure rate for extremely short intervals of time, or continuous time, is determined using the Hazard function h(t). This establishes a vital connection between proportional hazard models and survival analysis in determining the risk of developing DKD at time t, conditional on survival to that point:

$$h(t) = \lim_{\Delta t \to 0} \frac{[\{p(t \leq T < t + \Delta t\} / T \geq t]}{\Delta t} \quad \ldots\ldots (3)$$

Equation (3) displays the likelihood that a unit that endures until time t will experience the event in the next instant. The hazard can be expressed in terms of the survival function as:

$$h(t) = \frac{f(t)}{s(t)} \quad \ldots\ldots (4)$$

$$= -\frac{d}{dt} \ln S(t) \quad \ldots\ldots (5)$$

This represents the proportion of individuals contracting the event at time (t).

Time-to-event data may be analysed using one of three non-parametric approaches:

i. The Kaplan-Meier Method

According to Bradburn *et al.* (2003), this technique is utilized to plot survival curves for various patient cohorts. The Kaplan-Meier estimator for survival at time t_i is expressed as follows:

$$S(t_i) = \prod_{0 \leq j \leq n} Pr(T > t_i / T \geq t_i) \quad \ldots\ldots (6)$$

This can be further detailed as

$$S(t_i) = St_{j-1} \Pr(T > t_j / T \geq t_j)$$

$$= St_{j-1} \frac{n_j - m_j}{n_j}, \text{ for } 0 \leq t \leq T \quad \ldots\ldots (7)$$

Here j = 1, 2……n represents the total number of failure times recorded, T represents the failure time characterised by distribution F and density f, m_j is the number of failures that occur at time t_j and n_j is the number of individuals at risk at that same time t_j. The study will generate survival curves utilizing the Kaplan-Meier estimator. Staub & Alexandros (2011) provide the following general KM-formula for plotting the K-M curves:

$$S\,(t_j) = \prod_{i=1}^{j} Pr(T > t_i / T \geq t_i) = S\,(t_{j-1}).\, Pr(T > t_j \mid T \geq t_j) \text{…………..……....(8)}$$

In the absence of censoring, this simplifies to:

$$S(t_{(j)}) = \frac{surviving\ past\ t_j}{total\ number\ of\ people\ at\ the\ beginning} = \frac{R(t_{j+1})}{R\,(t(0))} \text{…………………..…....(9)}$$

ii. Life Table Method

Cutler and Ederer (1958) were the first to describe this method, commonly identified as the Actuarial or Cutler-Ederer method. This approach considers patients who are censored. Data are organized into tables and are grouped within intervals of fixed length. (Pagano, 2022).

iii. Nelson-Aalen Estimators

An approximation of the cumulative hazard rate which is non-parametric may be derived using the Nelson-Aalen estimator (Colosimo *et al.,* 2002). The fraction of the patient group present at time t that experiences an event per unit time is labelled as the instantaneous hazard. The cumulative hazard and survival are related as follows:

$$H(t) = \ln[S(t)] \text{ …………………………………………………………....(10)}$$

or

$$S(t) = e^{H(t)} \text{……………………………………….……………………..(11)}$$

According to Colosimo *et al.* (2002), the Nelson-Aalen estimator is defined as:

$$\widehat{A}(t) = \sum_{j:t_j \leq t} d_j / r_j, \quad (12)$$

Where r_j denotes the individuals at risk just prior to time t_j whereas d_j represents the deaths at that time. This formula, which calculates the ratio of deaths to those at risk, helps estimate the hazard at each specific death time, t_j. The expected number of deaths in the interval (0; t] per unit at risk is termed as the cumulative hazard up to the time t. It is effectively the summing up the hazards at all prior death times until time t. There is a compelling theoretical basis for this estimator in the theory of counting processes.

To determine the impact of selected independent variables on the survival rate of patients with DKD, the Cox proportional hazards model was employed. Each covariate is associated with a distinct coefficient, allowing their effects on survival rates to be evaluated through variation of these variables.

2.3.1 Log-Rank Test (LRT)

The null hypothesis states that there is no difference in the likelihood of an event (specifically, developing DKD) across populations at any specific time. To test this, the LRT is employed, which uses the dates of events, such as deaths (Bland & Altman 2004). The goal is to identify a function for which the distribution is either known or approximately known under the null hypothesis. A chi-square test statistic is calculated as described by Staub & Gekenidis (2011).

$$\chi^2 = \frac{\sum (O-E)^2}{E} \sim \quad (13)$$

where O represents observed total failures and E denotes the expected failures

The log rank test is particularly effective in detecting differences among cohorts when one consistently shows a greater risk of an event than the other. However, if the survival curves intersect—such as when comparing medical and surgical treatments—the test may not reveal any significant differences. Consequently, it is essential to plot the survival

curves when analysing survival data. The log rank test does not yield a confidence interval or an estimate of the difference's magnitude, as it serves solely as a significance test.

2.3.2 Cox Proportional Hazard Model

The model aims at concurrently assessing the impact that several variables have on survival. According to Johnson & Shih (2007) it is determined by

$$h(t/X) = h_0(\text{t})\ exp\,(X_1\,\beta_1 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + X_p\,\beta_p) \ldots\ldots\ldots\ldots \quad (14)$$

The predictors (covariates), $X_1, \ldots \ldots\ldots X_p$ are assumed to have an additive effect on log h(t/x). The current research the covariates which were; Age, Gender, Cardio-Vascular Disease, Hypertension, weight, marital status, level of education, family history of DKD, financial hardship, alcohol use and tobacco use. The log $h(t/X)$ which changes linearly with β_s, indicate that the impact of predictors is consistent across all time points t.

The hazard function h (t) denotes the instantaneous failure rate for a subject who has survived up to time t and may be defined as:

$$\text{h (t)} = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t \backslash T \geq t)}{\Delta t} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (15)$$

where Δt denotes an interval of time.

The hazard ratio for a patient who is characterized a batch of predictors X^* in comparison with another with a set of predictors X is given by

$$hr\ (X^*\ \text{X}) = \frac{\exp\,(X^*\beta)}{\exp\,(X\beta)} = \exp\,\{(X^*\text{- X})\,\beta\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (16)$$

The point estimate of the hazard ratio is expressed as;

$$\widehat{hr}(X^*\text{ : X}) = \frac{exp\,(X^*\widehat{\beta})}{\exp\,(X\widehat{\beta})} = \exp\,\{(X^*\text{- X})\widehat{\beta}\,\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (17)$$

Where $\widehat{\beta}$ serves as the maximum likelihood estimate of β.

The survival function, S (t), represents the probability that a subject will last longer than a specified time t. Theoretically, this probability decreases from 1 to 0 as time progresses from 0 to infinity. The survival function can be expressed in terms of the hazard function as:

$$S(t) = exp\left[-\int_0^t h(u)d(u)\right] \quad \text{(18)}$$

The hazard function can as well be defined as:

$$\text{h}(t) = \frac{d\text{S(t)/dt}}{\text{S(t)}} \quad \text{(19)}$$

This represents the percentage of subjects that die at time t. To interpret this, consider a person alive at time t. The probability of dying in a brief interval $[t, t + \Delta t)$ is given by:

$$q_t = \frac{\text{S(t)} - \text{S(t} + \Delta\text{t)}}{\text{S(t)}} \approx \text{h (t) } \Delta\text{t} \quad \text{(20)}$$

This can as well be expressed as;

$$\text{h(t)} = \frac{\text{No of deaths in [t; t} + \Delta\text{t)}}{\text{No of person} - \text{years at risk in [t, t} + \text{t)}} \quad \text{(21)}$$

Finally, H(t) = $\int_0^t h(u)d(u)$ represents the cumulative hazard risk.

In survival data modeling, the hazard function as well as the log-hazard is commonly used. With a constant hazard, h (t) $= \vartheta$ an exponential distribution of survival times has the density function given by:

$$p(t) = \vartheta e^{-\vartheta t} \quad \text{(22)}$$

2.3.2.1 Estimation of the Cox Proportional Hazard Model

The estimates for the parameters for this model are acquired by maximizing the partial likelihood. Let the observation of the i^{th}subject be represented as (x_{i1}, x_{i2}...,). Consider a conditional probability that a person dies at time t_j, one of the k recorded death times $\{t_1, t_2 \ldots, t_k\}$. Subjects are treated as responding independently, as expressed below:

$$L_j(\beta) = \frac{P(\text{individual with values } x_{(j)} \text{ dies at } t_{(j)})}{\sum_{k\epsilon R} t_{(j)} P(\text{ individual k dies at } t_{(j)})} \quad \text{.....(23)}$$

The intervals between consecutive time of death do not provide information about how covariates influence the hazard function, due to the arbitrary form of the baseline hazard. form. By substituting the probability of death at time t_j with that of death in the interval, $[t_{(j)}, t_{(j)}+\Delta)$, we can express:

$$L_j(\beta) = \frac{\frac{P(\text{individual with values } x_{(j)} \text{ dies at } [t_{(j)}, t_{(j)}+\Delta)}{\Delta}}{\frac{\sum_{k\epsilon R} t_{(j)} P(\text{ individual k dies at } [t_{(j)}, t_{(j)}+\Delta)}{\Delta}} \quad \text{.....(24)}$$

Contribution to the likelihood is;

$$\frac{h_j t_{(j)}}{\sum_{k\epsilon R} h\, k_{t(j)}} = \frac{h_0 t_{(j)} \exp x_j^T \beta}{\sum_{k\epsilon R} t_{(j)} h_o t_{(j)} \exp x_k^T} \quad \text{..... (25)}$$

Cancelling the baseline function in equation 20 reduces it to;

$$= \frac{exp\, x_j^T \beta}{\sum_{k\epsilon R} t_{(j)}\, exp\, x_k^T} \quad \text{.....(26)}$$

The joint partial likelihood is

$$l_p(\beta) = \prod_{j=1}^{n} \left[\frac{exp\, x_j^T \beta}{\sum_{k\epsilon R} t_{(j)}\, exp\, x_k^T}\right]^{\delta_j} \quad \text{.....(27)}$$

where δ_j is the censoring indicator, with a value of 1 indicating an event while a 0 indicating censoring. For uncensored observations the partial likelihood is given by:

$$\text{L } (\beta) = \prod_{Yi\ UC} \frac{\exp\,(X_{i\beta})}{\sum_{Y_{j\geq Y_i}} expX_j\beta} \quad \text{.....(28)}$$

The log partial likelihood is represented as:

$$l(\beta) = Log\ \text{L } (\beta) = \sum_{YiUC} \{ X_i\beta - \log [\sum_{Y_{j\geq Y_i}} expX_j\beta]\} \quad \text{..... (29)}$$

The joint likelihood is defined as;

$$\prod_{i=1}^{n} [f_i(t_{(i)})]^{\delta_i} [S_i(t_{(i)})]^{1-\delta_i} \quad \text{.....(30)}$$

This can be written as:

$$=\prod_{i=1}^{n} [h_i(t_{(i)}S_i(t_{(i)})]^{\delta_i}[S_i(t_{(i)})]^{1-\delta_i} \ldots\ldots (31)$$

or simply as:

$$=\prod_{i=1}^{n} [f_i(t_{(i)})]^{\delta_i}[S_i(t_{(i)})]^{1-\delta_i} \ldots\ldots (32)$$

This further expands to:

$$=\prod_{i=1}^{n} [\frac{exp\ x_{j\beta}^T}{\sum_{k\epsilon R} t_{(j)}\ exp\ x_k^T}]^{\delta_i} \times \prod_{i=1}^{n} [\sum_{k\in R(t_{(i)})} h_i(t_{(i)})]^{\delta_i} [S_i(t_{(i)})] \ldots\ldots (33)$$

2.3.2.2 Model Assumptions Checking Techniques

Testing the coefficient β is equivalent to verifying the assumption of the Cox proportional hazards model. This assumption may be assessed by employing several techniques, each of which has a recognized testing protocol and associated p-value.

a) Likelihood Ratio Test

This test can be applied within the proportional hazards model by hypothesis testing to evaluate the goodness of fit between the null and alternative models. For it to be valid, a single model must be nested within another, Such that it is the null hypothesis. In order to verify the proportional hazards assumption, Cox (1972) proposed including $X_2(t)$, a time-dependent variable, for the model. This variable can be the product of the variable of interest and time, denoted as $g(t)$. Considering $g(t) = t$, we have:

For the proportional hazards model:

$$L_{PHM} = h(t, X(t)) = h_0(t)exp(\beta_1 X_1 + \beta_2 X_2) \ldots\ldots (34)$$

For the extended Cox model:

$$L_{ECM} = h(t, X(t)) = h_0(t)\ exp(\beta_1 X_1 + \beta_2 (X_2 * t)) \ldots\ldots (35)$$

The test statistic is:

$$LR = -2\ log(L_{PHM}) - (-\ 2\ logL_{ECM}) \sim X_P^2 \ldots\ldots (36)$$

This test statistic follows a chi-squared distribution with p degrees of freedom, where p equals the difference in degrees of freedom between the models in Equations 35 and 36. The test statistics may also be considered as

$$D = -2\ln\frac{L(x_0 \mid t,\beta)}{L(x_1 \mid t,\beta)} \quad \text{(37)}$$

This can be written as:

$$= -2\ln \text{ (Likelihood of Null Model)} + 2\ln \text{ (Likelihood of Composite Model)}$$

b) **Schoenfeld Residuals Test**

The proportionate hazards assumption can be verified using partial residuals called Schoenfeld residuals. To derive the coefficients, the partial likelihood estimator is used. The residual for the i^{th} individual regarding the p^{th} variable is given by:

$$\hat{r}_{ip} = \delta_i\,(x_{ip} - \hat{x}_{wi,p}) \quad \text{where} \quad \hat{x}_{wi,p} = \frac{\sum_{j\in R(ti)} xjpe^{xj\beta p}}{\sum_{j\in R(ti)} e^{xj\beta p}} \quad \text{(38)}$$

The outcome, which is undefined for censored individuals, represents the difference between the covariate value of the individual experiencing the event at time t and the expected covariate value

c) Scaled Schoenfeld residuals test

Grambsch and Therneau (1994) recommended this method for testing time-dependent variables. It are defined as:

$$\hat{r}^*_{ip} = \frac{1}{[\widehat{var}(\hat{r}_{i1},\hat{r}_{i2},\ldots..\hat{r}_{ip})]}(\hat{r}_{ip}) \quad \text{(39)}$$

They also suggested that $E(\hat{r}^*_{ip}) + \hat{\beta}_p \approx \beta_p(t_i)$. Therefore, assuming that the proportional hazards assumption holds, $\hat{\beta}_p = \beta_p(t_i)$ and $E(\hat{r}^*_{ip}) = 0$, indicating that residuals exhibit a random walk over the time scale. Consequently, the null hypothesis states that there should be no regression coefficient between these residuals and a

function of time. Once calculated, the p-value for the connection between these residuals and time that is ranked can be determined.

d) Lin *et al*., (2006) Score Test

Lin et al. (2006), as referenced by Saegusa and Chen (2014), proposed "score-type" tests that utilize a natural smoothing spline representation for the non-parametric functions of time or covariates in assessing the proportional hazards assumption within the Cox model. This approach eliminates the need to explicitly define the function of time. The alternative model can be expressed as:

$$h\,(t \mid x_1\,, x_2) = h_0(t) exp\{X_1\beta + X_2\gamma(t) \quad \text{.....(40)}$$

Where, h_0 (t) is baseline hazard, X_2 is a scalar covariate of interest, X_2 is a scalar covariate of interest considered as a time-dependent variable, γ(t) as an arbitrary smooth function of time.

The smoothing spline $\gamma(t)$ is defined as:

$$\gamma(\text{t}) = \sum_{k=1}^{m} \delta_k\, \varphi_k(\text{t}) + \sum_{l=1}^{r} a_l \text{R}\,(\text{t}, t_l^0) \quad \text{..... (41)}$$

Where $m \geq 1$ is an integer, and $t^0 = (\,t_1^0, \ldots., t_r^0)$ is a vector of ordered events, with the assumption $0 < t_1^0 < \ldots. < t_r^0 < 1$. The constants δ_k and a_l represent coefficients.

The basis functions $\{\varphi_k(\text{t})\}_{k=1}^{m}$ correspond to the space of $(m-1)$th-order polynomials, and $R(t,s)$ is defined as:

$$R(t,s) = \int_0^1 \frac{(t-u)^{m-1} + (s-u)^{m-1}}{\{(m-1)!\}^2}\, du \quad \text{.....(42)}$$

The vector γ represents the values of γ(t) evaluated at each element of t^0 expressed as:

$$\gamma = \text{H}\delta + \sigma\text{a}$$

where

$$H=\begin{pmatrix} \varphi_1 t_1^0 & \cdots & \varphi_m t_1^0 \\ \vdots & \ddots & \vdots \\ \varphi_m t_r^0 & \cdots & \varphi_m t_r^0 \end{pmatrix} \qquad \sigma=\begin{pmatrix} R(t_1^0, t_1^0) & \cdots & R(t_1^0, t_r^0) \\ \vdots & \ddots & \vdots \\ R(t_r^0, t_1^0) & \cdots & R(t_r^0, t_r^0) \end{pmatrix}$$

e) Martingale-based Residuals Test

Martingale-based residuals were introduced by Therneau and Li (1999) as a method for assessing the proportional hazards model assumption. These martingale residuals represent the difference between the observed and expected number of events for the i^{th} subject at time t:

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_{ii}(u)\, exp\,(\beta' X_i) d\hat{h}_0(u) \text{……………….…………….(43)}$$

In this equation, $\widehat{M}_i$(t) is shorthand for $\widehat{M}_i(\infty)$, $N_i(t)$ indicates the number of events experienced by the i^{th} subject up to time t, and $\hat{h}_0$ is the cumulative baseline hazard. By plotting the cumulative martingale residuals against follow-up time, one may assess whether there is violation of the proportional hazard's assumption.

For all the methods mentioned above, a significant p-value ($p < 0.05$) indicates that the variable of interest is time-dependent and not constant over time, suggesting a violation of the proportional hazards' assumption.

2.3.2.3 Characteristics of the Cox Proportional Hazards Regression Model

a. The model is semi-parametric.
b. It does not impose specific assumptions about the form of $h(t)$ (the nonparametric aspect of the model).
c. It assumes a parametric form for the effect of predictors on the hazard.
d. The Cox proportional hazards model focuses more on estimating parameters than on determining the shape of the hazard function.

2.3.3 Support Vector Machine (SVM)

2.3.3.1 An Overview on Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) focused on developing algorithms that enable computers to learn from data and experiences without explicit programming. This capability allows computers to make predictions or draw conclusions based on data. In clinical research, ML is increasingly utilized for deep learning, pattern recognition, and predictive modeling (Vamathevan *et al.,* 2019). Machine learning can generally be categorized into two types based on the learning system's structure and the data available: supervised learning and unsupervised learning.

It is possible to learn classification rules from examples through supervised learning. An array of training examples, each identified as a member of a specific class, is fed into a learning algorithm. The goal of the algorithm is to generate a classification rule that will enable accurate assignment to these classes. The learning system's job is to generate a rule, or set of rules, that will enable precise data prediction (Shavlik & Dietterich, 1990). Common examples of supervised learning algorithms include support vector machines (SVM), decision trees, logistic regression, and linear regression.

In contrast, unsupervised learning employs unlabelled datasets for model training. It is up to the model to sift through the data on its own and identify trends and connections. Notable examples of unsupervised learning algorithms include Principal Component Analysis (PCA) and k-means clustering.

2.3.3.2 Support Vector Machine (SVM) Modelling

Given that Support Vector Machines (SVM) for survival analysis are less recognized compared to Cox regression, this section begins with a basic overview of the SVM concept. The research will adopt the same annotations used in the 'survivalsvm' R package publication by Fouodo *et al.* (2018).

According to Fouodo *et al.* (2018), an SVM is formulated with the assumption that the covariate $\boldsymbol{X} \in \mathbb{R}^d$ and the target variable $\boldsymbol{Y} \in \{-1,1\}$. When the two classes can be

separated linearly, there exists a linear function $f(\boldsymbol{x}) = \psi x + b$ such that $\boldsymbol{y} f(\boldsymbol{x}) > 0$. The objective of the SVM is to identify the separating hyperplane $H(\psi, b) = \{x | \langle x, \psi \rangle + b = 0\}$ that maximizes the margin between the two classes. The margin is defined as the minimum distance from any data point to the hyperplane.

The data points that lie at this margin distance are referred to as support vectors, as they define the margin and must satisfy either $f(\boldsymbol{x}) = 1$ or $f(\boldsymbol{x}) = -1$. In cases where the classes cannot be separated linearly, misclassifications may occur. This is managed by introducing slack variables $\xi_i \geq 0$, which allow misclassifications but also impose penalties. The slack variable for a specific instance i is calculated as $\xi_i = |y_i - f(\boldsymbol{x}_i)|$. Thus, $\xi_i = 0$ if i indicates correct classification, $\xi_i < 1$ means the point is correctly classified inside the margin, $\xi_i > 1$ denotes misclassification, and $\xi_i = 1$ indicates the point lies directly on the hyperplane.

To implement the SVM approach, the following optimization problem is established in primal space:

$$\begin{aligned} \min_{\psi, b, \xi} \quad & \frac{1}{2} \| \psi \|^2 + \gamma \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & -(\boldsymbol{y}_i(\langle \boldsymbol{x}_i, \psi \rangle + b) + \xi_i - 1) \leq 0 \end{aligned} \qquad \text{..... (44)}$$

$$\text{with } \xi_i \geq 0, i = 1, \ldots, n$$

In this formulation, ψ, b and the slack variables ξ_i are unknowns, n represents the number of instances, and $\gamma > 0$ is a regularization parameter that balances the maximum margin and the penalties for misclassification. Instead of solving this optimization directly in primal space, it is transformed into a dual problem, optimizing the Lagrange function in dual space (Bishop, 2007). The model is fine-tuned to identify optimal parameters. The formulation assumes linear separability of the two classes, which may not always hold true. Rusell & Norvig (2010) showed that a set of n data points can be separable in an $(n-1)$-dimensional space, indicating the potential need for a higher-dimensional space.

Three strategies have been proposed to address survival analysis using SVMs: regression (Shivaswamy *et al.,* 2007), ranking (Van Belle *et al.,* 2008), and a hybrid approach (Van Belle et al., 2011). The ranking method treats survival analysis as a classification problem with an ordinal target variable (Herbrich *et al.,* 1999), focusing on predicting risk ranks among individuals rather than survival times (Van Belle *et al.,* 2007). For two individuals i and j at time t, where an event occurs for j but not necessarily for i their predictions based on the SVM concept are $\langle \psi, F(x_i) \rangle + b$ and $\langle \psi, F(x_j) \rangle + b$, respectively. The ranking approach seeks to ensure that the ranking of $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, holds, meaning $(\boldsymbol{y}_i - \boldsymbol{y}_j)(\langle \psi, F(\boldsymbol{x}_i) \rangle - \langle \psi, F(\boldsymbol{x}_j) \rangle) > 0$. If $\boldsymbol{y}_j < \boldsymbol{y}_i$, then it must hold that $\langle \psi, F(\boldsymbol{x}_i) \rangle - \langle \psi, F(\boldsymbol{x}_j) \rangle > 0$. Thus, as proposed by Van Belle et al. (2007), the ranking method can be expressed as the optimization problem:

$$\min_{\psi,\xi} \frac{1}{2} \| \psi \|^2 + \gamma \sum_{\substack{j<i \\ \delta_i=1}} v_{ij} \xi_{ij} \quad \text{.....(45)}$$

subject to $\langle \psi, F(x_i) \rangle - \langle \psi, F(x_j) \rangle \geq 1 - \xi_{ij}$

and $\xi_{ij} \geq 0, i, j = 1, ..., n$

$$\text{where } v_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are comparable} \\ 0 & \text{else} \end{cases}$$

Observations i and j are deemed comparable if their survival times t_i and t_j satisfy $t_i < t_j$ and $\delta i = 1$ meaning the shorter observed time is not censored. Assuming the shortest survival time observation is uncensored, each observation $i = 2, ..., n$ will have at least one comparable counterpart. The problem presented here is equivalent to maximizing the concordance index (C-index) defined by Van Belle et al. (2007) over the comparable pairs for a given prediction function u:

$$CI_n(u) = \frac{1}{n(n-1)} \sum_{v_{ij}=1} I\left[\left(u(x_i) - u(x_j)\right)(t_i - t_j)\right] \text{.....(46)}$$

where $I(a) = 1$ if $a > 0$, and $I(a) = 0$, otherwise.

The regression approach is based on the support vector regression (SVR) (Vapnik, 1998) idea and aims at finding a function that estimates observed survival times as continuous outcome values y_i using covariates $\boldsymbol{x}_i$. The hybrid approach combines the regression and ranking approaches. In this work, the study implemented the survival SVR as proposed by Shivaswamy *et al.* (2007) and the hybrid approach (Ven Belle *et al.,* 2011). They are further discussed in section 3.8.2 in this document.

2.4 Comparative Analysis of Cox Regression and Support Vector Machine Models

The belief that machine learning is fundamentally more potent than conventional regression models has been the foundation for the growing use of these models in clinical research. However, a systematic review of empirical studies found that there were no appreciable differences in the model performance between logistic regression and machine learning models (Christodoulou *et al.,* 2019). The external generalizability of conclusions regarding the model performances is restricted by such comparisons.

There are a small number of simulation studies (Smith *et al.,* 2022) that compare machine learning and statistical methods for risk prediction for time-to-event data, but they frequently favour the novel approach. In order to predict time-to-event outcomes (the progression of diabetes to DKD), the study compares the effectiveness of the machine learning model Support Vector Machine (SVM) with Cox regression models under various data-analytic scenarios.

Calculating the C statistic for survival data, also known as "Harrell's C" or the "Concordance Index," is a popular method for assessing the predictive performance of models (Harrell et al., 1982; Ishwaran et al., 2008). The number of concordant pairs of observations divided by the total number of comparable pairs yields the C statistic. A prediction rule that is not informative has a value of C = 0.5, while perfect association is represented by a value of C = 1. This suggests that interpreting Harrell's C-index is simple.

Harrell's C was suggested by Ishwaran *et al*. (2008) as a means of assessing the predictive performance models. The Random Survival Forest (RSF) model of machine learning was suggested to use it in the study. Therefore, the Concordance index might be a useful metric to compare the SVM model and Cox regression. In previous biomedical applications, the C-index typically falls between 0.6 and 0.75. For instance, Van Belle *et al.* (2011) and Zhang *et al.* (2013) reported estimates in this range.

CHAPTER THREE
METHODOLOGY

3.1 Study Area

The study was conducted in Meru and Kirinyaga counties in Kenya, which were selected purposefully. In Meru County, the study was carried out at Meru Teaching and Referral Hospital, a regional referral facility in Meru town. The facility not only serves Meru county residents but also the neighbouring counties of Tharaka Nithi, Marsabit and Isiolo. The facility is situated 271.49 km from Nairobi at latitude 0.0506˚N and longitude 37.6534˚E on the northeast slopes of Mount_Kenya. In Kirinyaga County, Kerugoya Level 5 Hospital in Kerugoya town, was the study site. The facility is situated 115km from Nairobi, at latitude 0.5333˚S and longitude 37.2785˚ E on the slopes of Mount Kenya. The county is mainly inhabited by the Kikuyu community. These facilities were selected for this study because they are the largest public health care facilities in their respective counties and they are endowed with a state of art renal unit.

3.2 Study Design

A retrospective survey design was adopted to gather the information on patients with diabetes. This is because secondary data was used having been collected at a particular moment in time. The study sought to describe the variables in the population.

3.3 Study Population

Data for this study was obtained from diabetic patients who had attended sought treatment at Kerugoya Level 5 Hospital and Meru Teaching and Referral Hospital from January 2018 to July 2024. A total of 756 diabetic patients met the inclusion criteria. Among the diabetic patients, 396 were female and 360 were male.

3.4 Inclusion and Exclusion Criteria

To be eligible for inclusion in the study, the patient must have been a diabetic, aged 18 years and above, and have signed the informed consent. However, patients who were under 18 years old, whose medical report showed kidney issues before diabetes, those

critically ill and those who declined signing the informed consent were excluded from the study.

3.5 Instrument Validity and Reliability

3.5.1 Pilot Study

Pilot study to pre-test the research tool was be carried out at Chuka level 5 referral hospital, Tharaka Nithi County, which is of the same level as Meru Teaching and Referral and Kerugoya level 5 hospitals. Twelve diabetic patients were used. According to Bell *et al.* (2018), a minimum sample size of 12 respondents is applicable for a pilot study.

3.5.2 Validity Test of the Research Instrument

To determine the validity of the questionnaire (Appendix I), the university supervisors and the experts in the area of study carefully checked the questionnaire in order to improve its content validity. Necessary adjustments in the questionnaire were made so as to make it clear to the respondents for the purpose of meeting the objectives of the study hence enhancing the validity.

3.5.3 Reliability Test of the Research Instrument

In order to determine reliability, a pilot study was carried out utilizing the Test-Retest method of data assessment, which entailed giving the same instrument to the same group of subjects twice at two different times with a gap in time between the first and second test. To ascertain the correlation between the two set scores, Pearson's product moment correlation coefficient formula was employed. It is given by

$$r = \frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{[N\sum X^2-(\sum X)^2]\left[N\sum Y^2-(\sum Y)^2\right]}} \qquad (47)$$

Where, r is the Pearson’s correlation coefficient, N is the total number of pairs of test and retest scores, X and Y are the test and retest scores respectively. A coefficient of 0.70 or more indicated that the instrument was reliable to be used in a study (Tuckman, 1972).

3.6 Data Collection Procedures

Both primary and secondary data was collected for this study. The various participants were approached during their hospital visits for clinics. Upon agreeing to take part in the study, they read and signed the informed consent form. Primary data was collected using semi-structured questionnaires (Appendix I) on behavioural and demographic factors and linked to the secondary data obtained. Secondary data for the period 2018-2024 was obtained from the hospital records of both facilities on relevant medical history of the patients in the study. The data was collected using a checklist (Appendix II). The dependent variable in the study was the time taken to contract DKD after diabetes while there were 13 independent variables of interest including gender, age, weight, comorbidities such as cardiovascular disease and hypertension, behavioural factors such as physical activity, alcohol and tobacco use, and socio-demographic factors like marital status, level of education, family history of DKD, working status and financial difficulties faced by the diabetic patients. Financial difficulties were measured by their ability to meet payments of home and hospital bills, healthy feeding, and ability to acquire required medication. The boxes checked in the questionnaire enabled the researcher to access whether or not the patient was facing any financial hardships. These variables were then used for analysis in the study.

3.7 Data Analysis

Data exploratory was carried out using the excel package then exported to R software for analysis. The data was split into two; training set (70%) and test set (30%). String variables were appropriately converted into categorical variables. Survival analysis among relevant factors was first conducted using the non-parametric Kaplan-Meier method to plotting the Kaplan –Meier curves to understand differences in survival. Log rank tests were used to determine significance of different categorical covariates.

Both univariate and multivariable Cox regression analysis on the dataset was conducted. Univariate Cox regression helped analyse the risk of various variables in the data. The goal was to find the variables that are significantly associated with DKD. To do this, the

p-values of the log-hazard ratios of each variable were obtained. The covariates with p-values less than 0.05 were statistically significant in predicting DKD survival rate.

The R packages survminer and survival were extensively used in the study to create and validate the Cox regression models. In order to create survival prognostic models for diabetic patients, the survival Support Vector Regression method was chosen due to the unique characteristics of the sample size. The study used a number of R packages, including but not limited to survMetrics, survival, survminer, mlr, and survivalsvm, that are contained in R version 4.4.1 (Appendix III). To obtain the training set for model estimation and the test set for performance evaluation, pre-processing and partitioning of the data were carried out.

3.8 Statistical Model

3.8.1 Cox Proportional Hazards Model

In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. Cox proportional hazard model helped remove variables which had little or no impact on survival and eventually had a model, which only contained variables which affect DKD progression survival rate. The model formula is as follows

$$\log h_i(t) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots\ldots\ldots\ldots + \beta_{13} X_{i13} \text{ or equivalently}$$

$$h_i(t) = \exp(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots\ldots\ldots\ldots + \beta_{13} X_{i13}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (48)$$

where, i is a subscript for the observations, while the $X's$ are the co-variates (Hypertension-X_1, Gender -X_2, Age-X_3, Weight-X_4, Cardio-vascular disease-X_5, Marital status-X_6, Level of education-X_7, Family history of DKD-X_8, Financial hardship-X_9, Alcohol use-X_{10}, Tobacco use-X_{11}, Physical exercise-X_{12}, and Employment-X_{13}), βi is their respective coefficients and baseline hazard function $h_0(t)$. The constant α this model represents a kind of log-baseline hazard, since $\log h_i(t) = \alpha$ when all of the $X's$ are equal to zero.

The objective was to determine whether or not the addition of the various DKD covariates improved the model $h(t/x)$. Beginning with the null hypothesis $H_0 : \beta_1 = 0$, With composite hypothesis $H_1 : \beta_1 = 0$ at α=0.05. In essence, this is determining whether a suggested covariate significantly affects the model;

$$h(t/x_1) = h_0(t)e^{\beta_1 x_1} \quad \text{.....(49)}$$

This is done by using the test statistics below;

$$D = -2\ln\frac{L(x_0 \mid t,\beta)}{L(x_1 \mid t,\beta)} \quad \text{..... (50)}$$

Reject H_0 if $D > \chi^2$ at α = 0.05. Comparing the difference of one parameter in the model, there is 1 degree of freedom for the $\chi^2 >$ statistic, which is equal to 3.84 with 95% confidence. So, if D > 3.84 rejected H_0 , concluding that a certain covariate is an important factor in determining survival rate. D is the test statistic.

In order to compare survival curves based on hypothesis testing regarding the patients' survival rate, the Log-Rank test was employed. The different impacts that the DKD predictors had on the survival rate of diabetic patients were determined in large part by using Kaplan-Meier estimators. A plot of percentage survival (y-axis) against time (x-axis) was displayed by the Kaplan-Meier curves. In addition, the univariate Cox regression analysis for each variable was done. for both categorical and numeric variables. The Scaled Schoenfeld's residual tests helped evaluate the covariates. A significant relationship between residuals and time contradicts the proportional hazards assumption, while a non-significant relationship supports it.

3.8.2 Support Vector Machine (SVM)

In this study, modelling of the predictors and being able to predict the survival of diabetic patients before DKD is classification but a regression problem. Therefore, Support Vector

Machine for survival analysis (SVMs) were incorporated. The study therefore used the hybrid and the regression approaches in the study.

For the survival Support Vector Regression (SVR) the problem was formulated as follows:

$$\begin{aligned} \min_{\psi,b,\xi,\zeta_*} \quad & \frac{1}{2} \| \psi\|^2 + \gamma \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \textit{subject to} \quad & y_i - \langle \psi, F(x_i) \rangle - b \leq \xi_i, \\ & \delta_i(\langle \psi, F(x_i) \rangle + b - y_i) \leq \xi_i^* \\ \textit{and} \quad & \xi_i, \zeta_i^* \geq 0, \end{aligned} \quad \text{..... (51)}$$

where $i = 1, ..., n,$ is a subscript for the observations δ is the censoring indicator, n is the number of observations for the training set, in this case 530 individuals and F is a function that maps observed covariates to the feature space. The training data points are projected to be more easily separated by a hyperplane in the feature space, which is a higher dimensional space than the original space in which the data is observed. Since the feature space implies a higher dimensional space, the inner product $\langle \psi, F(\boldsymbol{x}_i) \rangle$ is calculated using a kernel function to reduce runtime (Vapnik, 1995). Assuming no censoring, i.e., $\delta_i = 1$, the inequality constraints are the same as in the classical SVR problem. However, if censoring occurs, the second constraint is reduced to $\xi_i^* \geq 0$.

The hybrid approach (Van Belle *et al.,* 2011) combines the regression and ranking approaches in the survival SVMs problem to form the optimization problem

$$\begin{aligned} \min_{\psi,b,\varepsilon_\xi,\xi^*} \quad & \frac{1}{2} \| \psi\|^2 + \gamma \sum_{i=1}^{n} \varepsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \langle \psi, F(\boldsymbol{x}_i) \rangle - \langle \psi, F(\boldsymbol{x}_{j(i)}) \rangle \geq \boldsymbol{y}_i - \boldsymbol{y}_{j(i)} - \varepsilon_i, \\ & y_i - \langle \psi, F(\boldsymbol{x}_i) \rangle - b \leq \xi_i, \\ & \delta_i(\langle \psi, F(\boldsymbol{x}_i) \rangle + b - \boldsymbol{y}_i) \leq \xi_i^* \\ \text{and} \quad & \varepsilon_i, \xi_i, \xi_i^* \geq 0 \end{aligned} \quad \text{.....(52)}$$

where $i = 1, ..., n$ is a subscript for the observations. The study took into account the corresponding Lagrange function in dual space and solved the quadratic optimization problem in order to solve the optimization problems.

3.8.3 Model Adequacy and Validation

Model validation was done by splitting my data into two sets, one for training and one for validation. The Cox model was tested for linearity and proportionality. The study used the concordance index (C-index) and log-rank P value as the metrics to evaluate model accuracy in order to compare the Cox regression and the Support vector machine for survival analysis (SVMs) models. For Cox Regression models, the area under the (ROC) curve (AUC) was further used to achieve an ideal situation

Harrell's C (C- Index) (Harrell et al., 1982) is given by

$$C = \frac{\sum_{i,j} I(\breve{T}_i > \breve{T}_j) \cdot I((\eta j > \eta i) \cdot \Delta j}{\sum_{i,j} I(\breve{T}_i > \breve{T}_j) \cdot \Delta j} \quad \text{.....(53)}$$

Where pairs of observations in the sample are denoted by the indices i and j. The number of concordant pairs of observations divided by the total number of comparable pairs yields the C statistic. When the smaller survival time is censored, i.e., $\Delta j = 0$, pairs of observations that are not comparable are discarded by multiplying by the factor Δj in Equation (53). The concordance probability $P(\eta j > \eta i \mid \breve{T}_i > \breve{T}_j)$ is calculated using Harrell's C method, which compares the rankings of two independent pairs of survival times $\breve{T}_i$, $\breve{T}_j$ and predictions ηi , ηj. The concordance probability assesses if there is a correlation between high values of ηi and low values of $\breve{T}_i$ or the opposite.

3.9 Modelling Process

3.9.1 Cox Proportional Hazard Model

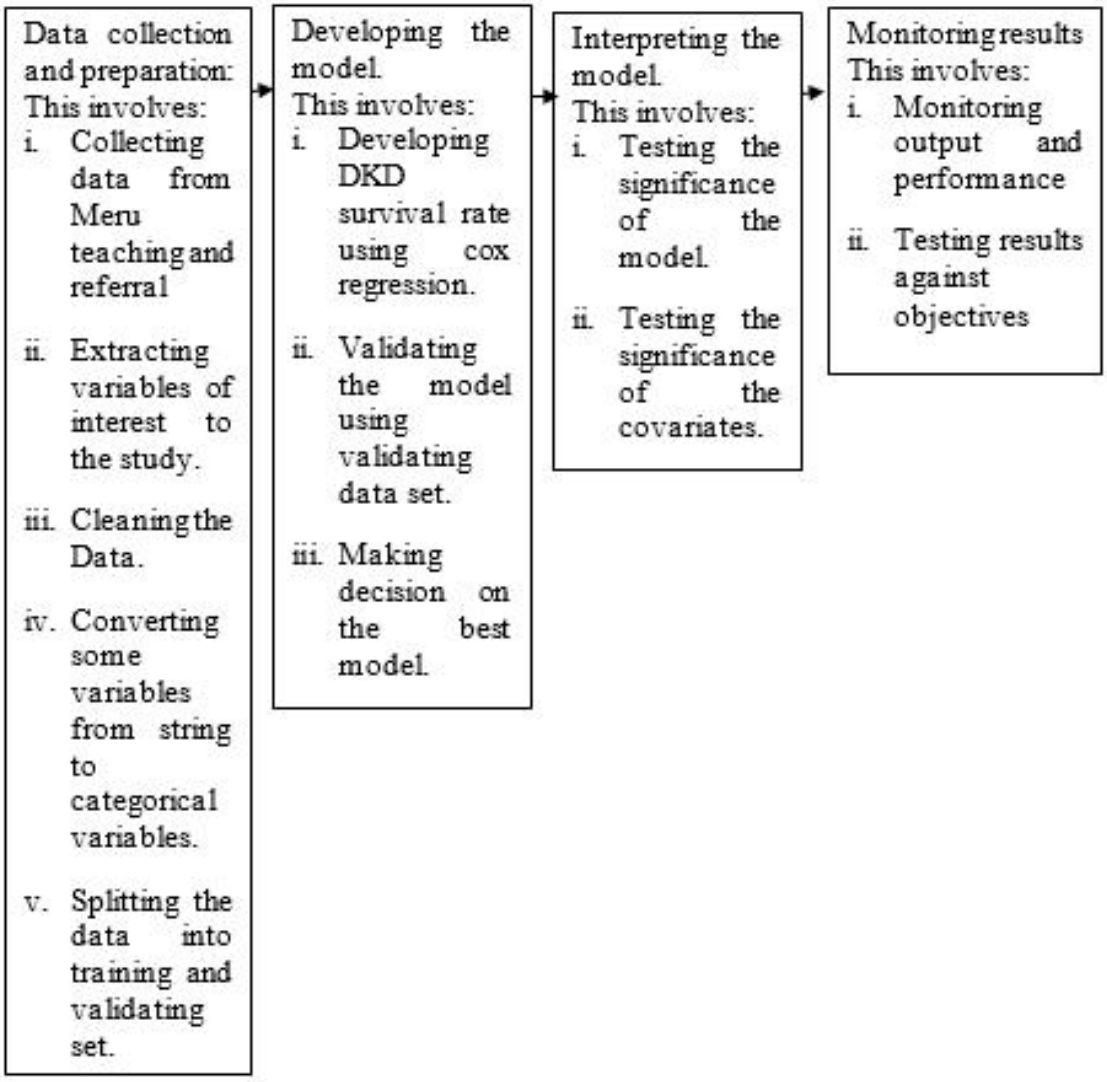

Figure 1: The Cox Proportional Hazards Modelling Process

3.9.2 Support Vector Machine Modelling Process

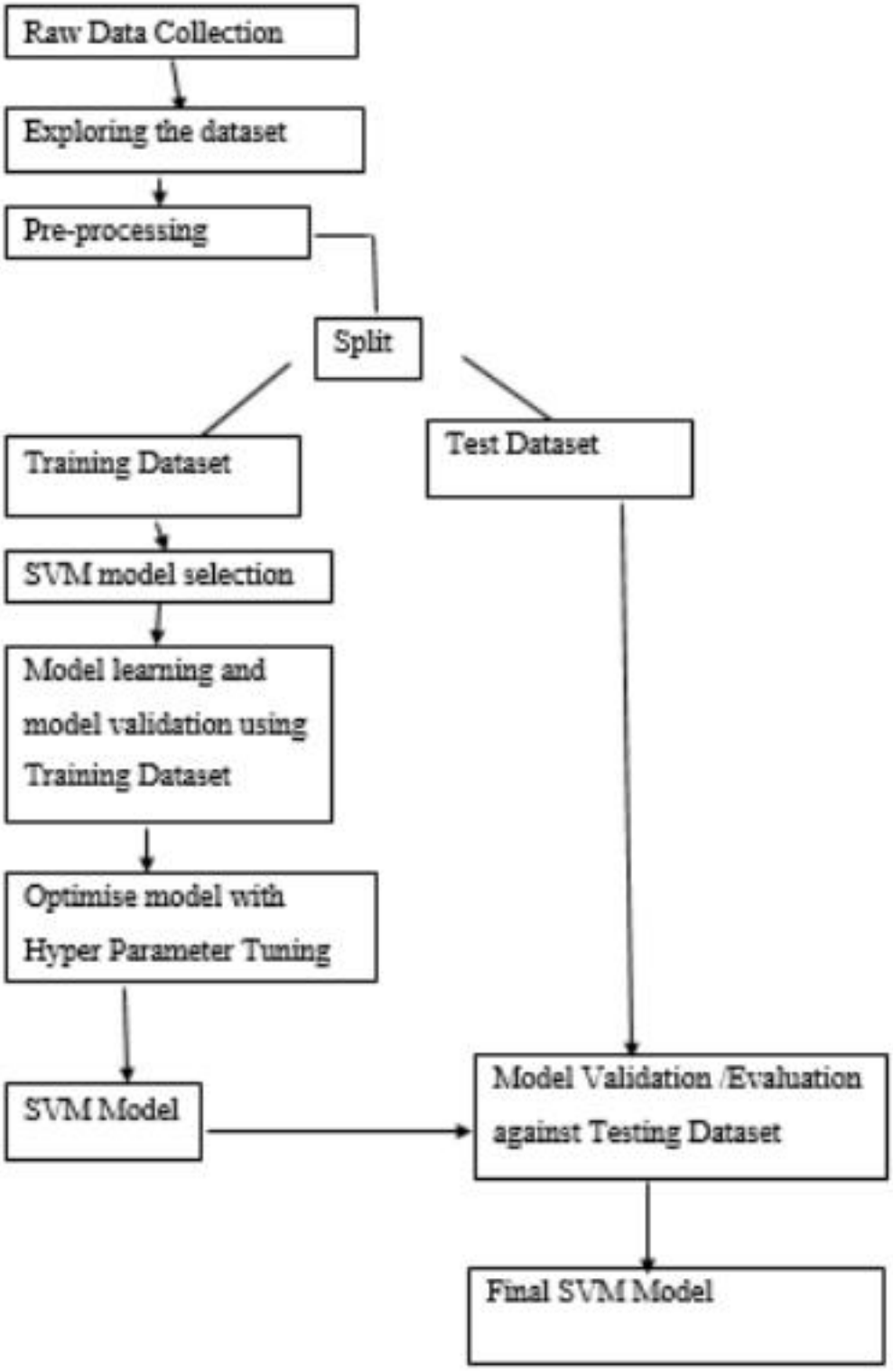

Figure 2: Diagrammatic scheme of support vector machine modelling process

Figure 2, represents the stages carried out in the study. After collection and exploring of the data, then it goes through pre-processing. Then the data is split into training and testing datasets, with 70% training dataset and 30% testing set. The training set is used when training algorithms and looking for suitable model while the test set is used to determine the performance of the model that has been produced. The next step involves modelling the predictors of DKD using support vector regression algorithms with hyperparameter optimisation. After the predictive modelling is generated it is validated using the test set and the performance of the model is validated using the performance metrics. Thus, coming up with the best model for the data.

3.10 Ethical Consideration

A research clearance was obtained from the Chuka University Ethics Committee. Then a research permit was obtained from National Commission for Science, Innovation and Technology (NACOSTI) before commencing of data collection (Appendix IV). Necessary clearance and permission were sought from the two hospitals involved in this study in order to obtain the data required. Strict adherence was made to the confidentiality and security of the acquired secondary data. All study participants signed informed consent after being fully informed about the purpose of the study and the necessity of their participation. The obtained data was analysed as outlined in the proposal. All literature cited were referenced appropriately. Data obtained was handled confidentially using coded names and not the actual patients' names.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1: Preliminary Analysis

The demographic, clinical and behavioural features of the patients were obtained from the 756 diabetic patients. Of the diabetic patients in the study, 286 (37.8%) had DKD while 470 (62.2%) patients were censored either for having withdrawn from the study before the occurrence of the event (DKD diagnosis), or had not suffered DKD by the time the study was completed. According to the current study, approximately four in every ten diabetic patients, had progressed to DKD. This finding shows a high prevalence for DKD among diabetic patients. The study agrees with that of Unamath & Lewis (2018) who showed that nearly 40% of all diabetic patients are affected by kidney disease. In addition, according to Selby & Taal (2020) DKD occurs in up to 50% of those living with diabetes and is a major cause of end-stage kidney disease (ESKD) that requires treatment with dialysis or renal transplantation and mortality among diabetic patients.

The findings of this study revealed that approximately 40% of diabetic patients have progressed to DKD, indicating a high prevalence of DKD amongst this population. This high prevalence of diabetes results to more complications that greatly affect patients' quality of life, increasing healthcare demand and escalating economic costs. Diabetic kidney disease (DKD) is a common complication among diabetic patients, driven by factors such as inadequate access to medical care, insufficient information about diabetes management, unhealthy lifestyles, and financial difficulties (Walker *et al*., 2021). These factors negatively impact glycaemic control and can lead to additional complications, such as diabetic nephropathy. Therefore, early detection and intervention for DKD are essential, especially given its potential severity. An appropriate prediction model(s) can help by identifying patients at higher risk of developing DKD, allowing for early intervention and targeted management strategies. Effective prediction models should incorporate demographic, clinical, and behavioural features to identify at-risk individuals and improve outcomes. Machine learning methodologies offer valuable insights and predictive capabilities by leveraging supervised learning algorithms, clustering techniques, and deep learning models to capture complex, non-linear relationships among

multiple risk factors (Esteva et al., 2019; Allen *et al*., 2022). Thus, machine learning approaches provide significant advantages over traditional statistical methods, enhancing the accuracy and reliability of DKD prediction.

4.1.1 Preliminary Analysis of the Numeric Variables in the Study

The mean time with diabetes for all the patients in the study was 12.14 years (SD = 6.91) with a minimum of 1 year and a maximum of 31 years. (Table 1). The skewness and the kurtosis values showed that the data satisfied the normality assumptions. The mean weight of the patients in the study was 84 kilograms, with a minimum of 56 kilograms and a maximum of 116 kilograms (Table 1). The study showed that the diabetic patients had a mean age of 41.29 years at the time of diabetic diagnosis with a standard deviation of 11.81 (Table 1). The minimum age at diabetic diagnosis was 15 years and a maximum of 76 years.

Table 1: Summary statistics for patients' characteristics (numeric)

Statistic	Time (years)	Patients age	Weight (Kgs)
Mean	12.14	41.29	83.98
Standard Deviation	6.91	11.81	10.82
Median	12	41	85
Range	30	61	60
Skewness	0.12	0.23	0.05
Kurtosis	-1.04	-0.58	-0.17
Maximum	31	76	116
Minimum	1	15	56

The findings of this study highlight the importance of weight loss in enhancing nephroprotection for diabetic patients, which can help to reduce the risk of developing DKD. High BMI is a known independent predictor of major renal events in patients with diabetes (Mohammedi *et al.* 2018). Accurate prediction of DKD risk using advanced methodologies, such as machine learning, can further enhance these efforts by identifying high-risk individuals early, allowing for timely interventions such as weight management and lifestyle modifications. Such predictive models can provide personalized recommendations, optimize patient care, and ultimately reduce the incidence and severity of DKD (Esteva *et al*., 2019; Topol, 2019; Nayak *et al*., 2024).

In this study, DKD patients had a slightly higher median age at diabetes diagnosis compared to the censored patients (Figure 3). This suggests that individuals diagnosed with diabetes at an older age are more likely to develop DKD. Therefore, adults over 40 years old, whether diabetic or not, are at risk of DKD, especially since many patients remain asymptomatic in the early stages of the disease. Pippitt *et al.,* (2016), recommended that screening for diabetes should be done for adults over 40 since they are at a higher risk of diabetes. The study's finding that patients with diabetic kidney disease (DKD) have a higher median age compared to those without DKD emphasizes the importance of early screening and monitoring for older adults. The observed correlation between age and DKD prevalence suggests the need for targeted preventive measures in this population to reduce the risk of complications from diabetes. To achieve this, it is recommended to use statistical tools such Cox proportional hazards models for analysing risk factors over time, and machine learning techniques for developing predictive models to identify high-risk individuals effectively.

A combination of statistical tool is critical since each of the statistical methodologies used in studies of DKD has its own limitations, which can affect the validity and applicability of the findings. For example, the Kaplan-Meier estimator assumes non-informative censoring, which may not always hold in clinical studies (Kleinbaum & Klein, 2012). The Cox proportional hazards model relies on the assumption of proportional hazards over time, which, if violated, can result in misleading conclusion (Bradburn *et al*., 2003). Logistic regression is sensitive to multicollinearity and assumes a linear relationship between the log odds of the dependent and independent variables (Hosmer *et al*., 2013). Propensity score matching requires careful selection of covariates and may not account for unmeasured confounding (Austin, 2011). Machine learning techniques such as random forests can be prone to overfitting and lack interpretability (Breiman, 2001; Aria *et al*., 2021), while Bayesian methods can be computationally demanding and sensitive to the choice of priors (Gelman *et al*., 2013; Crook *et al*., 2022).

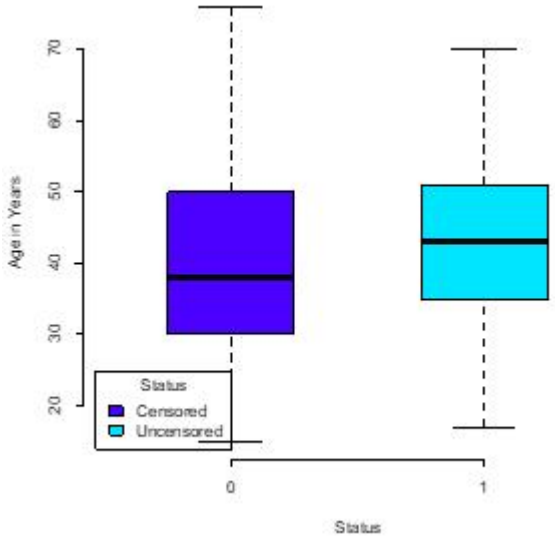

Figure 3: Diagrammatic summary for age

The DKD patients in the study had a duration of diabetes ranging from1 year to 25 years prior to developing DKD. Their median time before DKD was slightly above 15 years (Figure 4). However, the censored subjects in the study had lived to a maximum time of 31 years with a median time of less than 10 years.

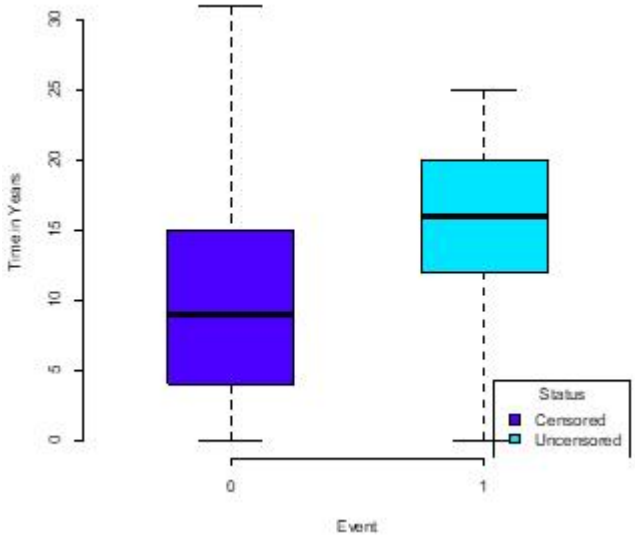

Figure 4: Summary on time since diabetes diagnosis

The median times in this study agree with a study by Varghese & Jialal (2023) that found out that patients with diabetes for a longer duration are more likely to develop DKD than patients with a shorter duration of diabetes. However, the fact that a patient in the current study was able to live at least 30 years without DKD probes the need to investigate further on the predictors of DKD in order to apply the findings on diabetic patients with the aim of alleviating incidence DKD. The long survival time may be attributed to advancements in drugs and therapies for diabetes complications. The observation that patients with a longer duration of diabetes are more likely to develop DKD suggests that the duration of diabetes should be a key predictor in any model developed to forecast the risk of DKD. The fact that some patients can live for over 30 years without developing DKD indicates significant heterogeneity in risk among diabetic patients. This suggests that there may be other protective factors or unknown predictors influencing DKD development. Predictive models must account for this heterogeneity by including a broad range of potential predictors, such as genetic factors, lifestyle variables and comorbidities. Machine learning techniques like random forests or gradient boosting can help identify complex interactions and non-linear relationships between these factors (Breiman, 2001; Friedman, 2001; Aghaabbasi and Chalermpong, 2023). Distiller (2014) concluded that absence of DKD after 15 - 20 years of diabetes appears to be a marker of long-term survival without DKD, suggesting the need for stratification in predictive models.

The median weight of patients with DKD was higher than that of patients without DKD (Figure 5). Increased weight is associated with worse glucose levels which may cause kidney damage. Anderson *et al*. (2005) attributed uncontrollable glucose and cardiovascular disease to obesity. Weight management could be crucial for improving diabetes and DKD outcomes. According to Centre for Disease Control (CDC), (2004), almost all adults with diabetes are overweight; more than half are obese. Given that the median weight of patients with DKD is higher than those without DKD, weight should be included as a critical predictor in any model aimed at assessing DKD risk. Predictive models should account for body mass index (BMI) or other measures of adiposity to better estimate the risk of DKD, as weight is strongly linked to glucose control and cardiovascular health, both of which are significant contributors to kidney damage

(Anderson *et al*., 2005). Since weight is a modifiable risk factor, models should consider not just the static weight of a patient but also changes in weight over time. Time-dependent covariates could be used to capture weight fluctuations and their impact on DKD risk. This approach allows models to dynamically assess risk based on current weight trends and interventions that aim at weight management. Predictive models should also account for the complex relationship between weight, glucose levels and DKD. Increased weight can lead to poor glucose control, which in turn exacerbates kidney damage (Kumar *et al*., 2023). Machine learning techniques, like random forests or gradient boosting, could be valuable for identifying non-linear interactions (More & Wolkersdorfer, 2023) between weight, glucose levels, and other factors contributing to DKD.

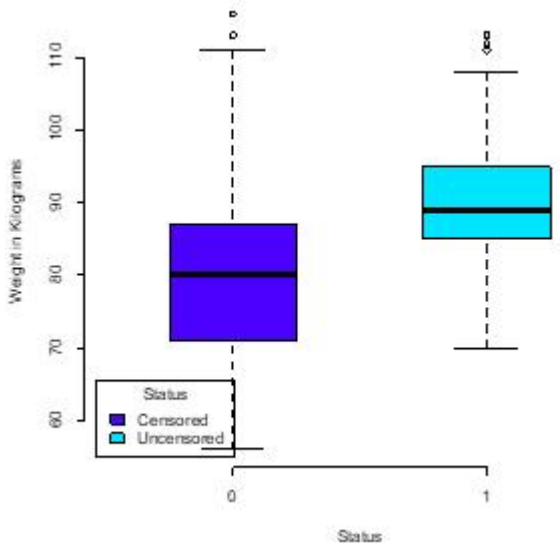

Figure 5: Diagrammatic summary for factor weight

4.1.2 Preliminary Analysis of the Categorical Variables in the Study

The study was composed of 396 (52.4%) female diabetic patients and 360 (47.6%) male diabetic patients (Table 2). More than half of the diabetic patients were also suffering from hypertension (486 patients, 64.3%). Comorbid cardiovascular diseases had affected 206 of the patient population translating to 27.2%. Among the participants of the study, 33.3% had received education up to primary school level, 36.8% had received education

up to secondary school level and 29.9% had received education up to tertiary level, translating to 252, 278 and 226 patients, respectively. About 51% of the patients had no spouses, either because they were widowed, separated, divorced or never married. Most of the patients were in a form of employment (56.9%) while the others were unemployed (8.7%) or were retired (34.4%) either due to medical reasons or old age. Accordingly, 56.1% of the participants agreed that they were facing some financial hardship. In the study, 494 patients (65.3%) had not used tobacco, and 498 (65.9%) did not have any alcoholism history.

Table 2: Summary statistics of patients' characteristics (categorical)

Feature		Frequency	(%)
Gender	Female	396	52.4%
	Male	360	47.6%
Hypertension	No	270	35.7%
	Yes	486	64.3%
Cardiovascular Diseases	No	550	72.8%
	Yes	206	27.2%
Education	Primary	252	33.3%
	Secondary	278	36.8%
	Tertiary	226	29.9%
Marital status	No Spouse	388	51.3%
	Spouse	368	46.7%
Tobacco use	No	494	65.3%
	Yes	262	34.7%
Alcohol use	No	498	65.9%
	Yes	258	34.1%
History of CKD	No	472	62.4%
	Yes	284	37.6%
Physical Exercises	Frequently	444	58.7%
	Rarely	312	41.3%
Financial Hardship	No	424	56.1%
	Yes	332	43.9%
Employment	Employed	430	56.9%
	Retired	260	34.4%
	Unemployed	66	8.7%

In this study, a high prevalence of hypertension and a substantial proportion of patients with a history of CKD highlight some of the health concerns in this patient population. According to studies, hypertension and diabetes are closely related and are key factors

that cause progression to DKD (Buffet & Ricchetti, 2012; Kim *et al*., 2022). The majority of patients engage in physical exercise frequently and are non-users of tobacco and alcohol, which could be relevant for their overall health profile. The data on education, financial hardship, and employment provides context for understanding the socioeconomic background of the patients, demographic, health and lifestyle characteristics of the patient population, which is crucial for tailoring health interventions and support services. Currently, Kenya's unemployment rate is 5.6% (O'Neill, 2024). As seen in the study, the rate of unemployment among the diabetic patients in the study population was higher by 3.1% without including patients who have attained retirement age of 60 years as per Kenya public service regulations. This may aggravate the challenge of financial hardships among the diabetic patients and thus increase prevalence of DKD.

The demographic and health-related characteristics of the study population have several implications for developing predictive models for DKD. To account for the interplay between these factors, predictive models may consider using multivariate analysis techniques that allow for the inclusion of multiple predictors and their interactions. Models may consider stratifying the risk predictions based on demographic and lifestyle factors to tailor interventions more effectively. Machine learning models can incorporate a wide range of features (Gachoki *et al*, 2022a), including demographic variables (e.g., age, gender, education level), comorbid conditions (e.g., hypertension, cardiovascular disease), lifestyle factors (e.g., tobacco use, alcohol consumption), and socioeconomic status (e.g., financial hardship). Machine learning algorithms can handle complex interactions between features (Gachoki *et al*, 2022b). For example, models can explore how the combination of high weight, poor glucose control, and financial hardship interacts to affect DKD risk.

4.2 Incidence of Diabetic Kidney Disease

Analysis using the Kaplan Meier survival function curve and the survival table showed that two diabetic patients contracted DKD within a year after diagnosis (Table 3). The percentage of the patients who survived without DKD within the third year was 98.7%. By the ninth year, 6.7% of the patients had suffered DKD and 93.4% had not. Only

11.1% of the patients had remained by the 31st year. Two patients were at risk of contracting DKD by the end of the 31st year. Generally, the Kaplan-Meier survival curve and survival table reveal a sharp decline in the proportion of patients remaining free of DKD over time, with significant survival reduction by the 31st year. The analysis shows a sharp decline in the proportion of patients remaining free of DKD as time progresses, with a notable reduction in survival by the 31st year, reflecting the long-term challenges and increasing risk of DKD and other health issues. The findings of the current study are in agreement with the study by Varghese & Jialal (2023) in that most DKD occurrences were registered 12 to 24 years after diabetes (Table 3). According to Varghese & Jialal (2023), while patients with type 2 diabetes mellitus may exhibit albuminuria at the time the diabetes is detected, DKD may typically develop 15 to 20 years later in individual with type 1 diabetes.

Table 3: Survival table for diabetic patients before diabetic kidney disease (DKD)

Time	No. at Risk	No. of DKD occurrences	Survival	Survival SE
0	756	2	1.000	0.00000
3	682	8	0.987	0.00422
6	592	8	0.974	0.00609
9	494	22	0.934	0.01023
12	394	42	0.842	0.01638
15	300	42	0.734	0.02108
18	208	44	0.607	0.02468
21	96	76	0.329	0.02734
24	30	38	0.138	0.02349
27	6	4	0.111	0.02250
30	2	0	0.111	0.02250

The mean survival time for diabetic patients before developing DKD in the current study was 19.1 years. The Kaplan-Meier curve indicates that the median survival time is 20 years (Figure 6). This survival time is higher than that of 16 years that was realised by Tekalign *et al.* (2023) in a study done in Ethiopia. The current study highlights that diabetic patients are at increasing risk of developing DKD as time progresses. Therefore, early intervention and continuous monitoring are crucial, as the probability of remaining free of DKD diminishes significantly over time. The high rate of DKD occurrence in later

months underscores the need for timely management strategies to improve long-term outcomes for diabetic patients.

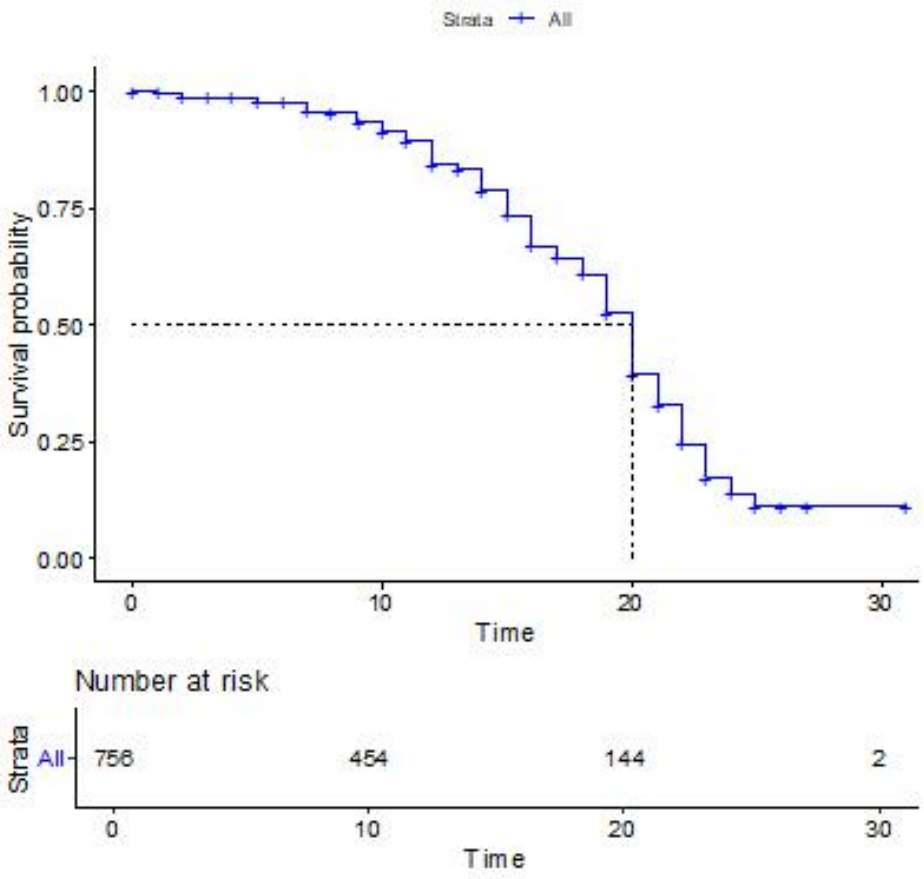

Figure 6: Kaplan Meier curve showing overall survival probability of diabetic patients

4.3 Identifying the Predictors of Diabetic Kidney Disease

Log-rank tests were performed to test for the presence of any significant differences among categorical variables survival curves. Further, Kaplan-Meier curves were plotted to visually attest the proportional hazards assumption of Cox regression. The following hypothesis was tested,

H_o: There is no difference in survival between the groups, against

H_A: The groups are statistically different at $\alpha = 0.05$.

Further, univariate Cox regression was performed to determine its independent effect on risk of DKD. The results of the Log Rank tests, Kaplan- Meier curves of the various

categorical variables, and univariate Cox regression analysis are outlined in the subsequent section.

4.3.1 Log- Rank Test and Kaplan- Meier Curves Results

The findings of this study revealed that there was a significant difference (p = 0.037) in time to contracting DKD between male and female patients. The female patients had a slightly longer median survival time (20 years) compared to male patients (19 years) (Table 4). The mean time for female diabetic patients of 19.64 years was also longer than the mean time for males of 18.37 years. On plotting Kaplan – Meier curves, they revealed a small margin between the genders, with female patients surviving longer than the male patients before DKD (Figure 7). The curves intersected at some points, but this did not affect the proportionality assumption (Figure 7).

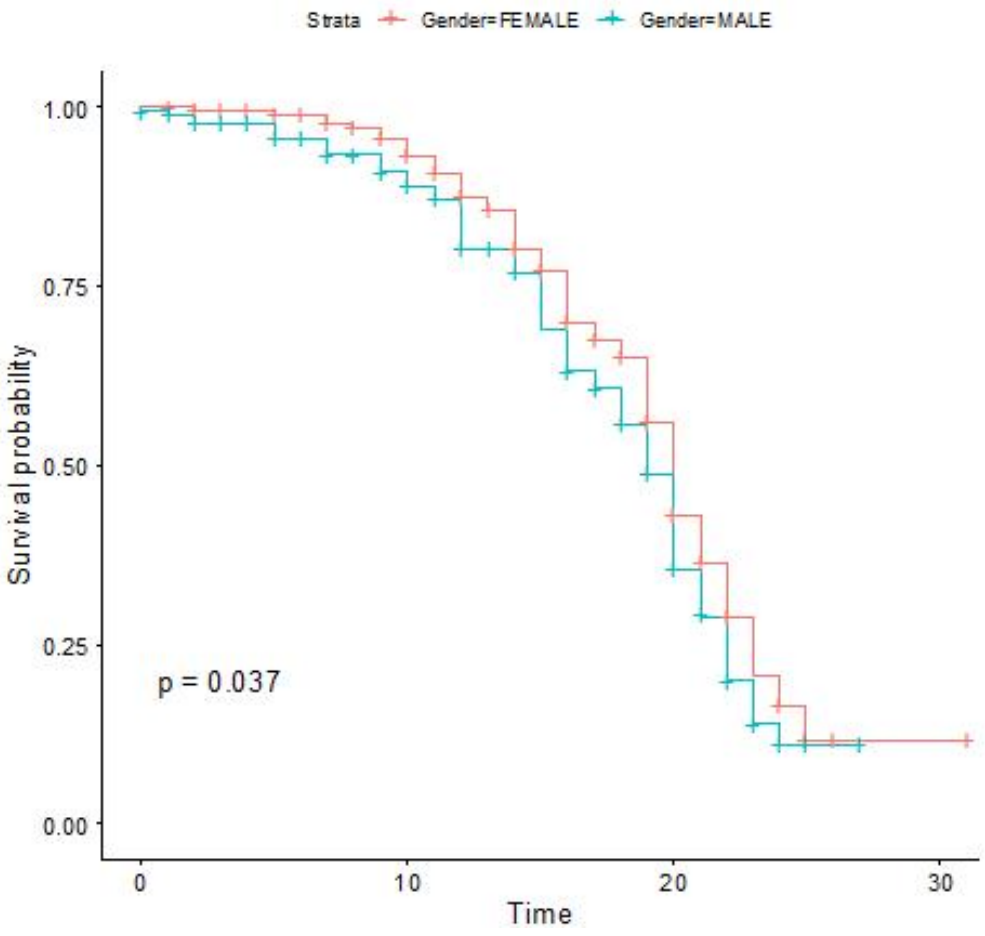

Figure 7: Kaplan Meier curves for gender

It was also observed that the hypertension significantly ($p < 0.05$) affects survival time, with patients without hypertension having a much longer mean survival time (28.65 years) compared to those with hypertension (18 years). This difference was statistically significant ($p < 0.0001$; Table 4). The median time for those without hypertension was undefined because the survival curve did not drop to or below 0.5 (Figure 8). There was a consistent drop in survival probability for the patients who were suffering from comorbid hypertension as compared to the fairly stable Kaplan - Meier curve for those without hypertension. The study findings demonstrate that hypertension significantly impacts survival time in diabetic patients. Recognizing the significant effect of hypertension on survival highlights the need for effective management strategies. Interventions may include lifestyle modifications, medication adherence, and regular monitoring to control blood pressure and improve survival outcomes (Muntner & Judd, 2017). Therefore, incorporating hypertension as a variable in predictive models may improves their accuracy by accounting for its significant impact on survival. Predictive models that include hypertension can provide better estimates of survival probabilities and guide targeted interventions to manage hypertension and its impact on DKD.

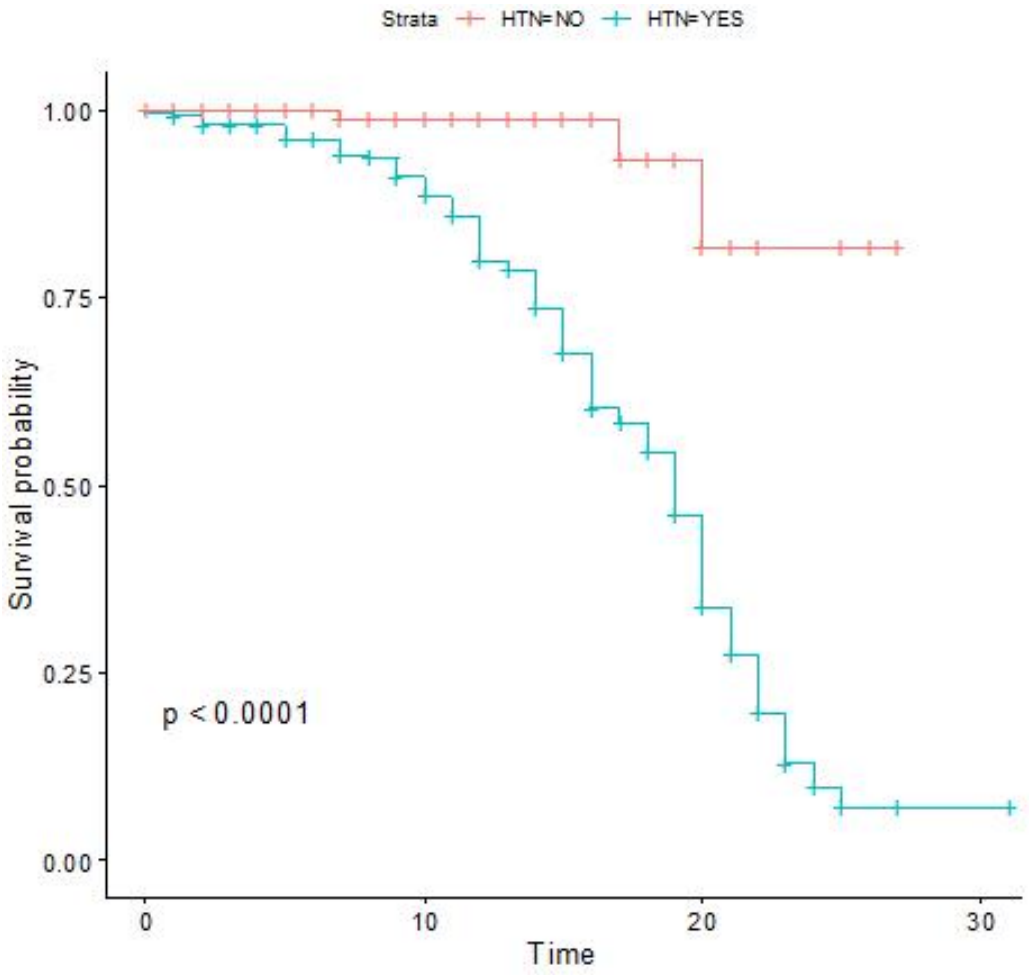

Figure 8: Kaplan Meier curves for hypertension

The presence of cardiovascular diseases significantly ($p < 0.05$) reduces survival time compared to those without cardiovascular diseases, with a more pronounced effect ($p < 0.0001$). Patients with cardiovascular diseases had a median time of 16 years, which was significantly shorter than for the diabetic patients without cardiovascular diseases (21 years; Table 4). The Kaplan - Meier curves for patients with cardiovascular diseases and those without were also significantly different with diabetic patients without cardiovascular disease having a higher survival probability, as depicted in Figure 9. This study highlights the critical role of cardiovascular health in predicting survival and managing DKD. Cardiovascular complications contribute to a faster deterioration of kidney function, impacting overall survival (Vart & Li, 2018). Incorporating

cardiovascular disease into risk assessment models offers a more holistic view of patient health.

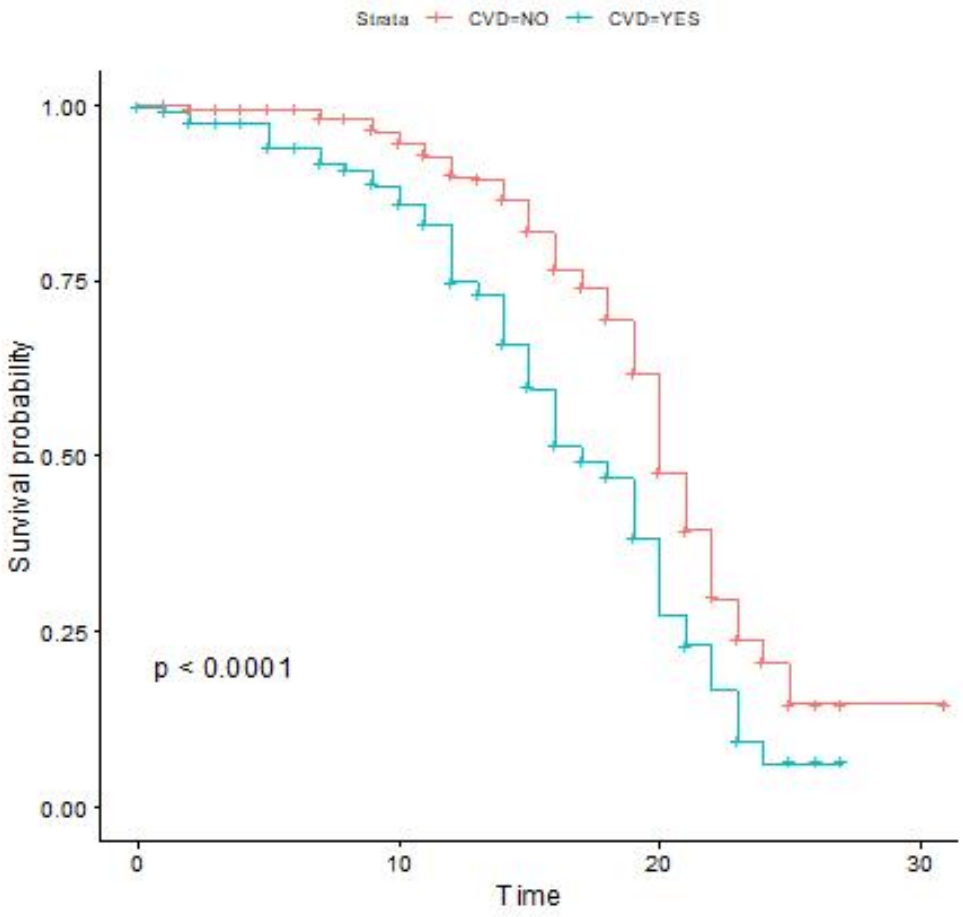

Figure 9: Kaplan Meier curves for cardiovascular diseases

The findings of this study showed that higher education levels are associated with longer period before DKD, with tertiary education showing the longest median survival time of 22 years (Table 4). The K – M curves showed that the curves were distinct for at least 15 years with the tertiary level having higher survival probability. Later on, the curves intersected and the survival probability did not seem to differ (Figure 10). However, the log rank test confirmed that they were statistically ($p < 0.05$) distinct. This suggests that including education level as a predictor in the model could improve its accuracy in estimating survival outcomes. Higher education levels often correlate with better health literacy, leading to more effective management of diabetes and associated complications.

Educated individuals are more likely to understand their condition, adhere to treatment plans, and engage in preventive health behaviours (Cutler & Lleras-Muney, 2010).

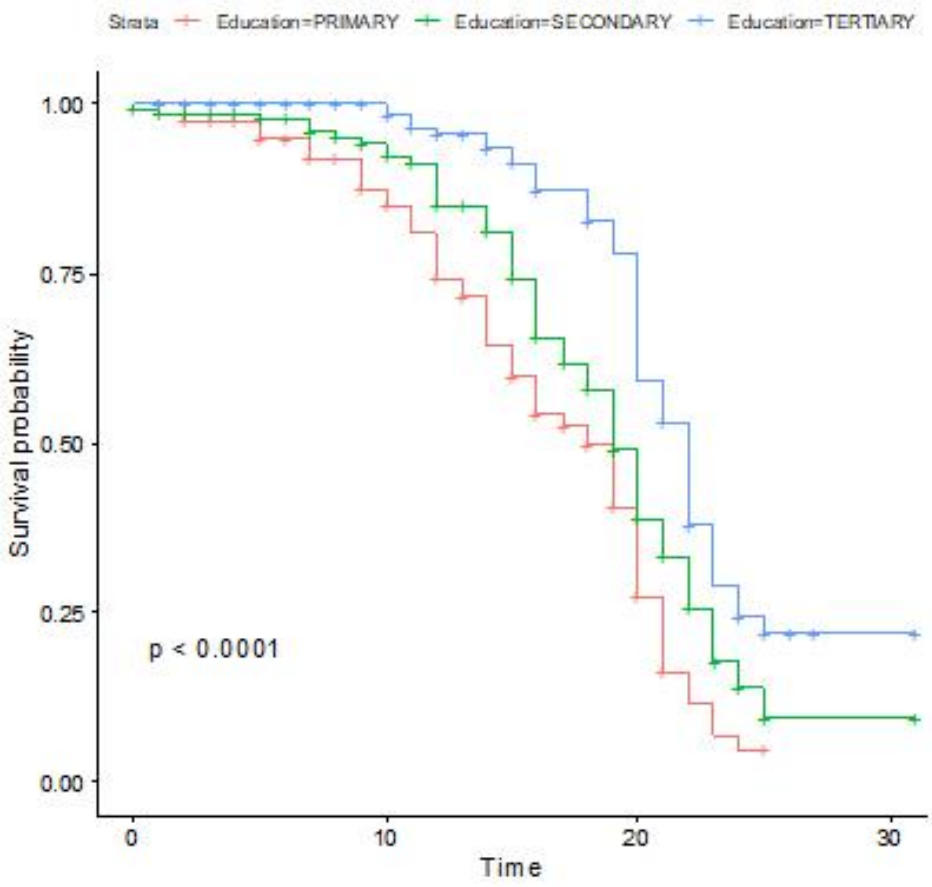

Figure 10: Kaplan Meier curves for level of education

Being married (spouse) was also associated with longer survival compared to those without a spouse (Table 4). Married diabetic patients had a longer 3 years overall DKD-free survival time than those without spouses (18 verses 21years, $p < 0.0001$). The K-M curve clearly showed that patients with spouses had a higher survival probability at each time instant compared to those without spouses (Figure 11). Married individuals tend to have longer survival times compared to their unmarried counterparts. This association may be attributed to various factors, such as emotional support, shared responsibilities, and better access to healthcare (Stenholm, & Head, 2018). This finding underscores the relevance of marital status in predictive models for DKD and survival outcomes. Models

that include marital status can identify patients who may benefit from additional support or interventions based on their social circumstances.

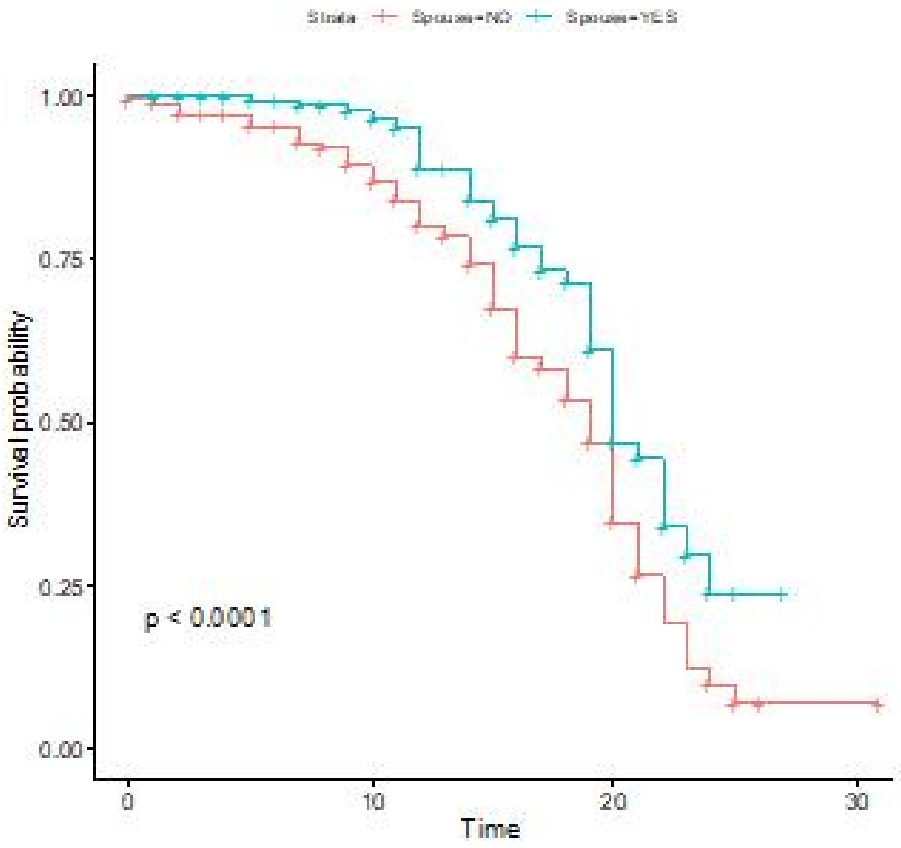

Figure 11: Kaplan Meier curves for marital status

The current results revealed that tobacco use significantly ($p < 0.05$) decreases survival time, with non-users having a much longer median survival time before developing DKD ($p < 0.0001$). The median survival time for non-users was 24 years as compared to the 17 years for the users (Table 4). These findings align with previous clinical studies indicating that smoking is a significant factor in the progression of diabetic kidney disease (DKD) in diabetic patients (Orth, 2000). Figure 12 presents Kaplan-Meier curves related to tobacco use, showing significant differences; at every point, non-smokers exhibit a higher survival probability than smokers. Smoking has a multifaceted harmful effect on renal physiology (Gündoğdu & Anaforoğlu, 2022) and is recognized as a key risk factor for DKD. This study highlights that tobacco use is an important predictor of

DKD. Research indicates that smokers face a greater risk of developing DKD earlier than non-smokers, with a notably accelerated progression of kidney disease among tobacco users, adversely affecting their overall survival (Stengel & Pueyo, 2017; Fiore & Jaén, 2020). Therefore, predictive models that include tobacco use can provide more nuanced risk assessments and support decision-making in clinical practice, leading to more effective management strategies. Models that consider tobacco use can outline patients who would gain the most from smoking cessation programs. These interventions can potentially improve survival rates and delay the onset of DKD.

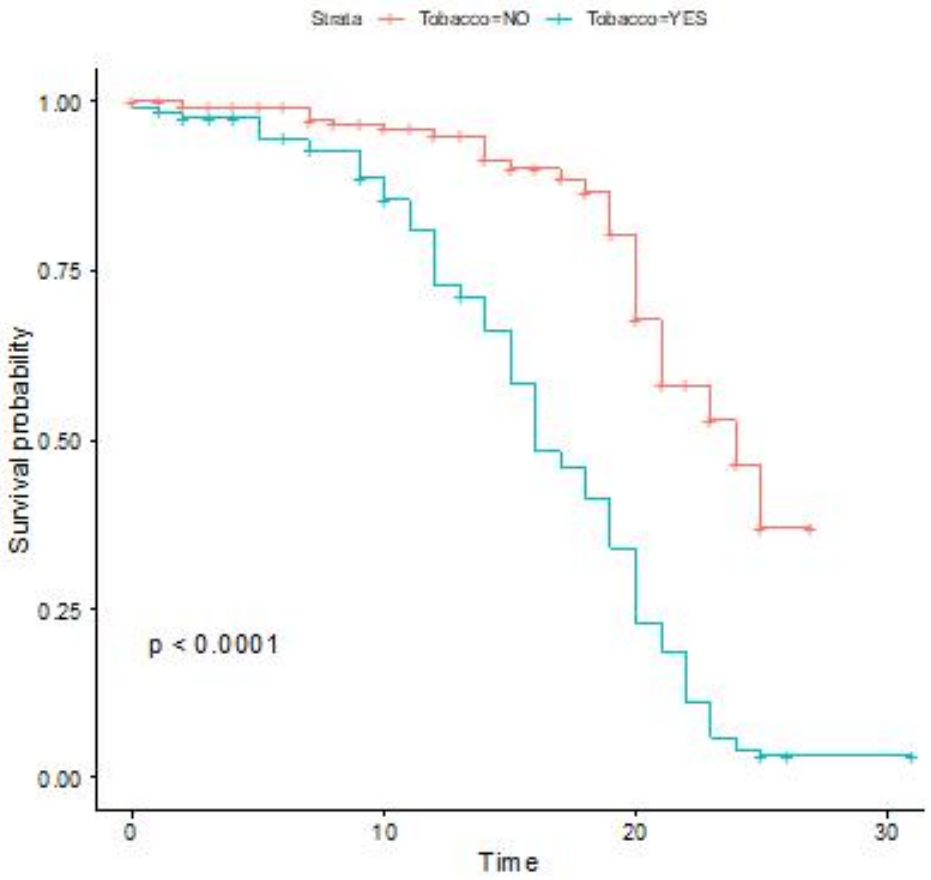

Figure 12: Kaplan Meier curves for use of tobacco

The median survival time for patients using alcohol was 18 years and that for patients who had abstained from alcohol use was 22 years (Table 4). Patients who did not take alcohol were able to survive longer without DKD. This was also confirmed with the K – M curves for alcohol use (Figure 13). The log-rank test indicated a significant difference between the curves ($p < 0.05$). Thus, the Kaplan-Meier (K-M) curves and the log-rank

test both support the finding that there is a significant difference in survival times between patients who did and did not consume alcohol. This indicate the importance of including this predictor in predictive modelling.

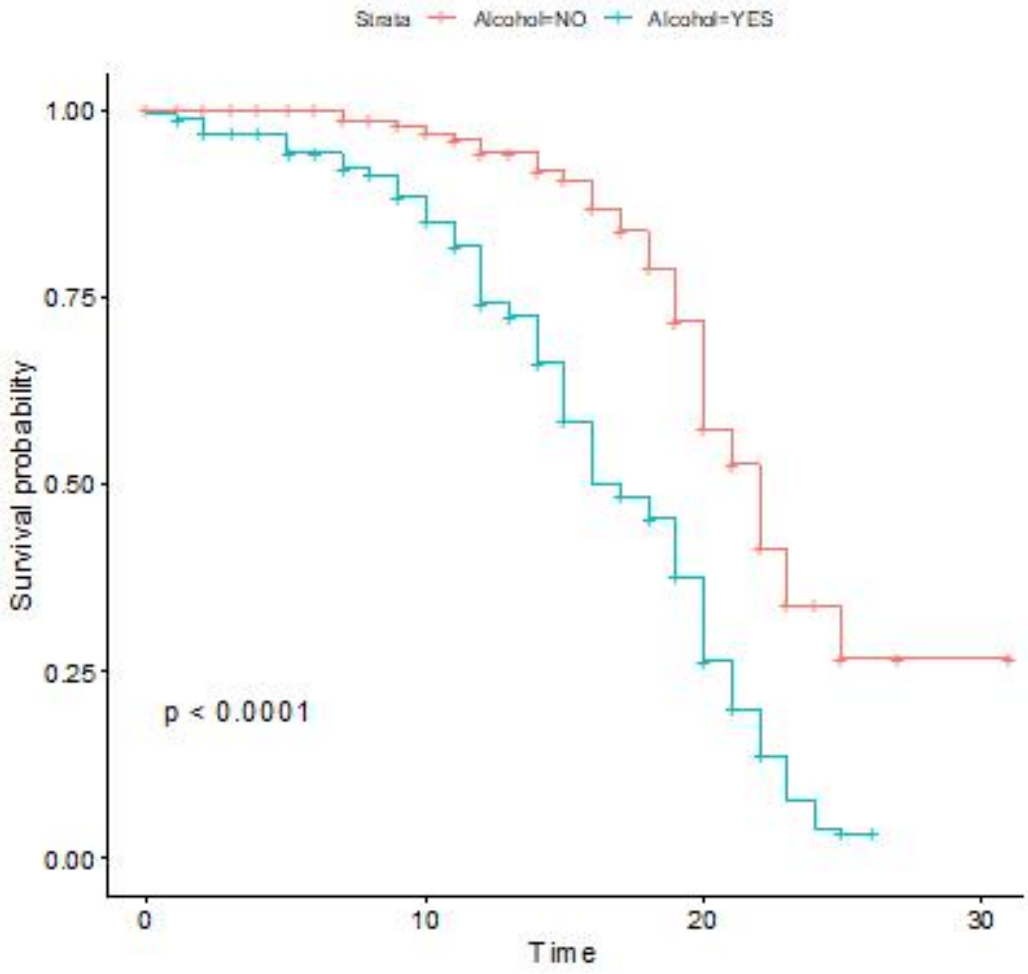

Figure 13: Kaplan Meier curves for use of alcohol

A family history of chronic kidney disease (CKD) significantly reduces survival time, with those without a history of CKD having longer survival. The mean survival for diabetic patients with a history of DKD was 16.32 years and 26.88 years for patients who had no history (Table 4). The log rank tests revealed a significant difference in the survival curves. This finding shows that having a family history of CKD can significantly increase risk of DKD. Figure 14 shows the two curves for the patients with or without a family history of kidney disease. The curves show that individuals without a history of CKD have a higher survival probability. A family history of CKD indicates a genetic

predisposition to kidney disease (Kovesdy & Liew, 2018). This genetic risk factor can exacerbate the progression of CKD in diabetic patients, influencing their overall survival. Incorporating family history of CKD into predictive models had been found to enhance their accuracy (Williams & Hall, 2021). This variable helps identify patients at higher risk of developing CKD and experiencing reduced survival times. Predictive models that account for family history of CKD can lead to personalized care plans.

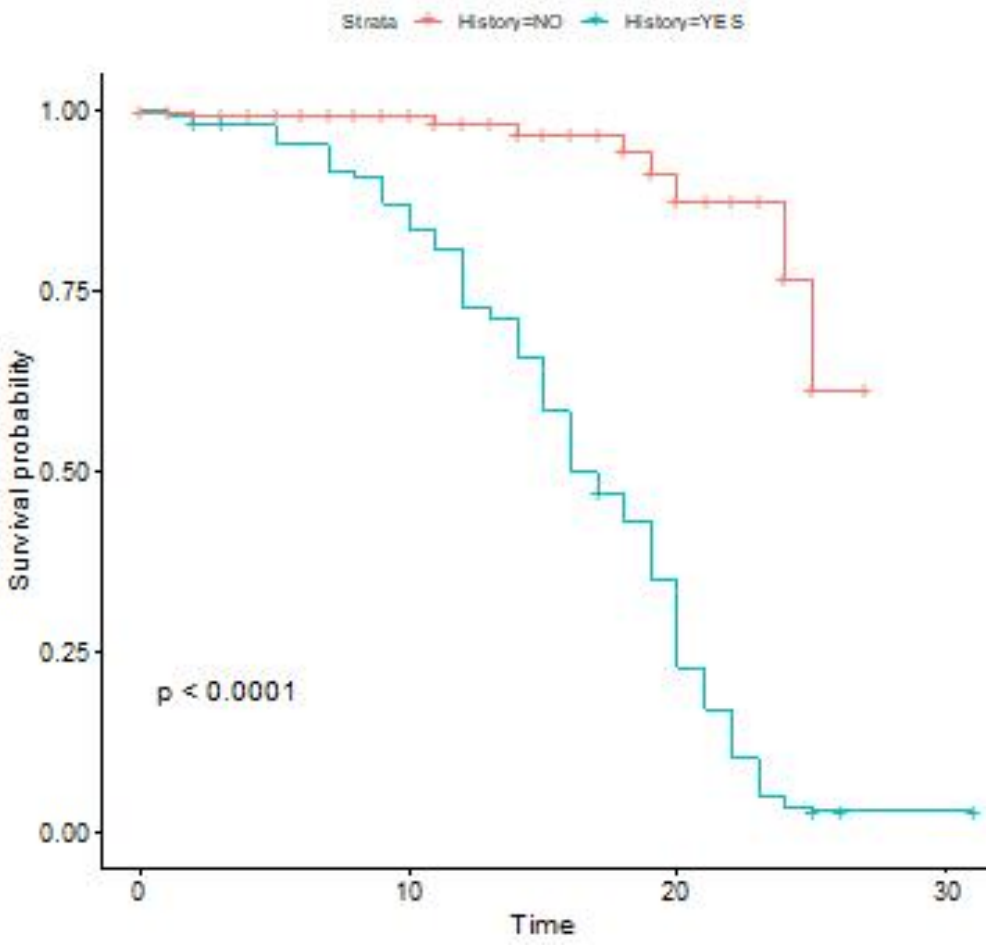

Figure 14: Kaplan Meier curves for family history of CKD

Regular physical exercise was associated with longer survival compared to rarely exercising with median survival times 22 years and 18 years, respectively (Table 4). The Kaplan - Meier curves vividly illustrate a significant disparity in survival rates between patients who engage in frequent exercise and those who did not (Figure 15). This visual

representation suggests a potentially crucial role of physical activity in prolonging the period before the onset of DKD. Similarly, the log rank tests confirmed a significant difference ($p < 0.0001$) in survival between patients who engaged in frequent exercise and those who did not (Table 4). This finding reveals that there is need for exercising frequently as it can lower a patient's hazards of getting DKD. This finding highlights the potential benefits of regular physical activity in managing diabetes and preventing complications. Several benefits accrue from regular physical activity, which includes improved glycemic control (Colberg *et al*., 2016), reduction in cardiovascular risk (Pescatello *et al*., 2014), enhanced renal health (Kovesdy, & Liew 2018) and weight management (Thomas *et al*., 2006). Therefore, integrating physical activity into predictive modelling for diabetes management offers the potential to enhance the accuracy of risk assessments, personalize treatment plans, and improve overall patient outcomes. By considering the various benefits of exercise models can better predict complications and support more informed healthcare decisions.

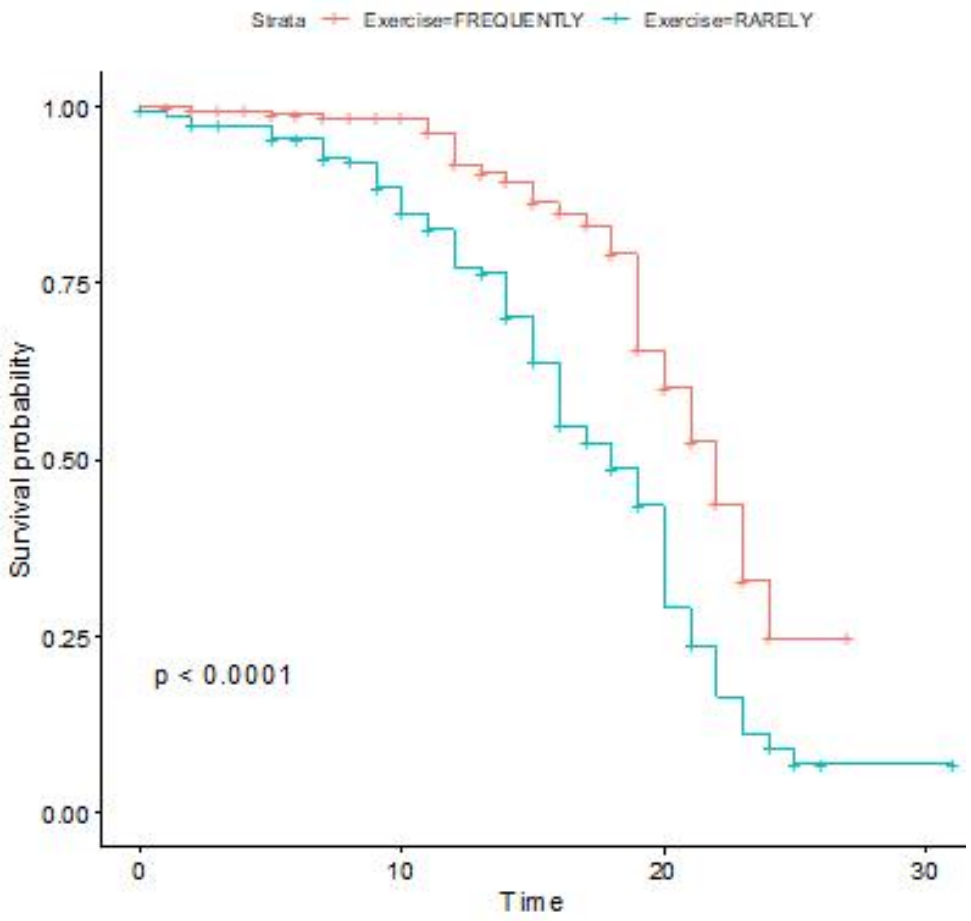

Figure 15: Kaplan Meier curves for physical exercises

Absence of financial hardship was associated with longer median survival (23 years) with those experiencing financial hardship having a shorter median survival time (18 years). Adequate financial support for diabetic patients would reduce their hazard rate of acquiring DKD. The K – M curves also were different (Figure 16). Performed log-rank tests ($p < 0.0001$) confirmed that there was a significance difference between the two survival functions Table 4). This finding shows the critical role that financial stability plays in the management and outcomes of diabetes. Financial hardship can impact access to healthcare services, medications, and necessary treatments (Gibbons & Kelsey, 2022). Limited financial resources may lead to delayed or inadequate care, which can negatively affect disease progression and survival. Hence financial hardship is an important predictor of DKD, influencing survival rates and disease progression.

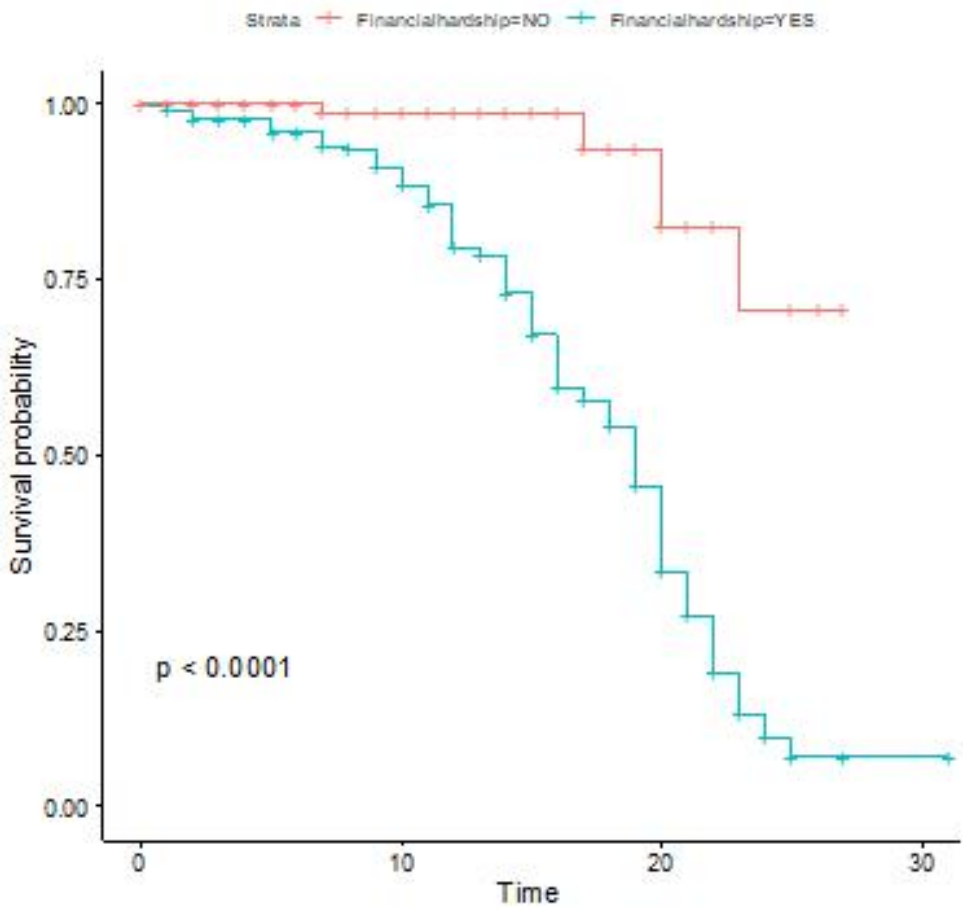

Figure 16: Kaplan Meier curves for financial hardships

The findings of the study also found out that employment status influences survival, with employed individuals (23 years) showing longer median survival compared to those retired (20 years) or unemployed (18 years) (Table 4). Therefore, being engaged in some form of employment or career had a positive impact on survival rate of diabetic patients before DKD. The constructed survival curves revealed some interaction specifically for the employed and the retired patients' survival curves in the first 15 years (Figure 17). However, later on there was a distinct difference in the survival curves. Log rank tests confirmed that indeed there is a statistical difference ($p < 0.0001$) in the three survival functions for the employed, unemployed and retired patient in the study (Table 4). This association suggests that being employed may contribute positively to survival, possibly because of better access to healthcare, social interactions, and financial stability (Schoenbaum & McCarty, 2019). Employment often provides access to health insurance,

regular medical care, and other benefits that can contribute to better management of chronic conditions and longer survival. Employed individuals may also experience less stress related to financial instability compared to retirees (De Graft-Johnson & Schneider, 2020). Incorporating employment status into predictive models enhances their accuracy by accounting for socioeconomic factors that affect health outcomes. This allows for a more comprehensive understanding of survival risks and helps in identifying individuals who might benefit from additional support. Predictive models that include employment status can provide better insights into survival probabilities and guide interventions tailored to the needs of employed, unemployed and retired individuals.

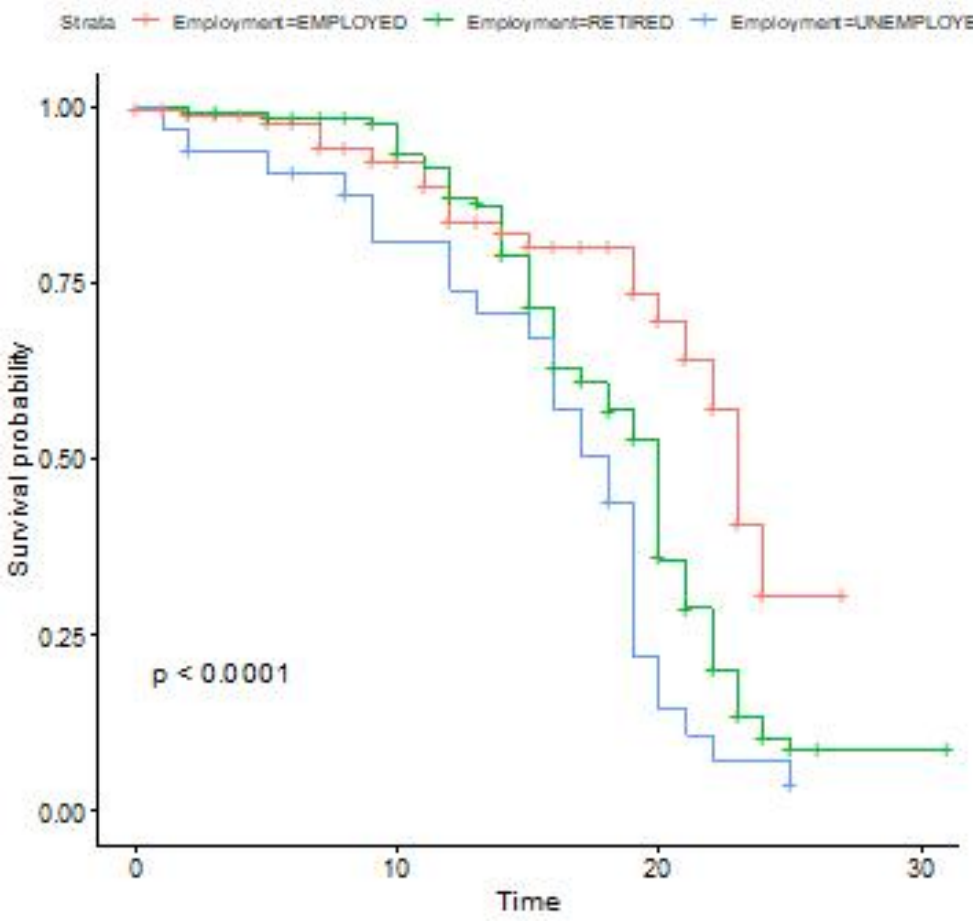

Figure 17: Kaplan Meier curves for employment

Table 4: Summary results from the log- rank tests

Variable	Category	Events	Median Time	Mean Time (95% C.I.)	χ^2 Test statistic	Log Rank's P-Value
Gender	Male	144	19	18.37	4.4	0.04
	Female	142	20	19.64		
Hypertension	No	6	NA	28.65	54.6	1e-13
	Yes	280	19	18.02		
Cardio-Vascular	No	150	21	20.97	30.4	4e-08
	Yes	136	16	15.39		
Level of Education	Primary	109	18	16.78	39.8	2e-09
	Secondary	128	19	18.95		
	Tertiary	49	22	22.19		
Marital Status	Spouse	206	20	21.08	20.3	7e-06
	No Spouse	80	19	17.81		
Use of Tobacco	No	56	24	23.64	111	<2e-16
	Yes	230	16	16.39		
Use of Alcohol	No	78	22	22.33	76.9	< 2e-16
	Yes	208	17	16.58		
History of CKD	No	18	NA	26.89	164	<2e- 16
	Yes	268	16	16.32		
Exercise	Frequently	64	22	21.94	44	3e-11
	Rarely	222	18	17.34		
Financial Hardship	No	81	23	27.12	56.6	5e-14
	Yes	205	18	17.90		
Employment	Employed	56	23	22.12	35.2	2e-08
	Retired	174	20	18.84		
	Unemployed	56	18	16.03		

4.3.2 Univariate Cox Regression Analysis

Further, a univariate Cox regression analysis was performed for each factor in the study and results were recorded (Table 5). Age was significantly associated with increased hazard (HR = 1.019; 95% confidence interval [CI], 1.007 - 1.031; p = 0.00184). This finding showed that increase in age by a year increased the hazards of getting DKD by 1.9% all other variables held constant. Age was statistically significant ($p < 0.05$) for the model. The univariate effect of gender was as well statistically significant (HR Male = 1.275; 95% CI [1.011 - 1.608]; p = 0.0401). Being of the male gender increased the hazard of DKD by 27.5%. Suffering from hypertension (HR = 10.951, CI [4.871 - 24.63]; $p < 0.0001$) was as well strongly associated with increased prevalence DKD in diabetic

patients holding other factors constant. From these finding, a patient who is hypertensive increase their hazard of DKD by a factor 10.951. A univariate Cox regression with comorbid cardiovascular disease (CVD) as the factor of interest discovered that having a comorbid CVD increased hazards of DKD in diabetic patients (HR = 1.8830; CI [1.493 - 2.375]; $p < 0.0001$).

Weight also showed a significant positive association (HR = 1.046; CI [1.035 - 1.056]; $p < 0.0001$) to the hazard ratio of contracting DKD. All other factors constant, an increase in weight of a diabetic patient would increase their hazards of progressing to DKD by 4.6%. Secondary level of education (HR = 0.652; CI [0.5024 – 0.845]; $p = 0.00124$) and tertiary level of education (HR = 0.352; CI [0.2495 – 0.495]; $p < 0.0001$) negatively affected the hazard rate as found in the current study (Table 5). From these findings, achieving secondary or tertiary education reduced the hazard rate of DKD by 34.2% or 64.8% respectively as compared to a patient whose highest level was primary education holding other factors constant.

Marital status indicated a protective effect for those with a spouse (HR = 0.562; CI [0.4336 – 0.7279]; $p < 0.0001$) in the univariate analysis. Patients who had spouses had decreased hazard by 43.8% compared to patients who did not have spouses (Table 5). Tobacco use (HR = 4.225; CI [3.152 – 5.665]; $p < 0.0001$) was found to significantly increase the hazard. Alcohol use (HR = 3.002; CI [2.315 – 3.897]; $p < 0.0001$) significantly increased hazard. According to this univariate analysis on alcohol, holding other factors constant, the intake of alcohol by a diabetic patient would increase their hazards of progressing to DKD by a factor 3.002. Diabetic patients who consume alcohol have a higher risk of DKD.

Having a family history of CKD caused patients to be more susceptible to DKD compared to those without such a history (HR=11.675; CI [7.235-18.84]; $p < 0.0001$). Patients who were not were rarely involved in physical exercises increased their hazards of contracting DKD by a factor 2.471 (HR = 2.471, $p < 0.0001$). Financial hardship also significantly increased hazard (HR = 9.934; CI [4.686-21.06]; $p < 0.0001$). Employment

status showed that both retired (HR = 1.845; CI [1.361 – 2.502]; p < 0.0001) and unemployed individuals (HR = 2.993; [2.062 – 4.347]; p < 0.0001) faced higher hazards when compared to employed diabetic patients.

Table 5: Summary results from univariate Cox regression analysis

Factor		Unadjusted HR exp(coef)	Lower 95%	Upper 95%	P-value
Age		1.019	1.007	1.031	0.00184
Gender	Male	1.275	1.011	1.608	0.0401
Hypertension	Yes	10.9525	4.871	24.63	7.05e-09
CVD	Yes	1.8830	1.493	2.375	9.2e-08
Weight		1.0457	1.035	1.056	<2e-16
Education	Secondary	0.6516	0.5024	0.845	0.00124
		0.3515	0.2495	0.495	2.19e-09
Marital Status	Yes	0.5618	0.4336	0.7279	1.28e-05
Tobacco	Yes	4.225	3.152	5.665	<2e-16
Alcohol	Yes	3.002	2.312	3.897	<2e-16
History of CKD	Yes	11.675	7.235	18.84	<2e-16
Exercise	Rarely	2.471	1.867	3.269	2.42e-10
Financial Hardship	Yes	9.935	4.686	21.06	2.12e-09
Employment	Retired	1.845	1.361	2.502	7.97e-05
	Unemployed	2.994	2.062	4.347	8.24e-09

4.4 Developing a Predictive Model for Diabetic Kidney Disease

4.4.1: Cox Proportional Hazards Model

The multivariable Cox proportional hazards model was utilized to assess the effects of various variables on the hazards of DKD in diabetic patients (Table 6). Age at diagnosis showed a significant positive impact on the hazard ratio (HR = 1.0486; CI [1.0296 - 1.0680]; p = 3.92e-07). Additionally, being male was linked to a significant increase in the hazard of DKD (HR = 1.3625; CI [1.0506 – 1.7671]; p = 0.0197). Hypertension however, produced statistically insignificant results (YES_ HR = 0.7617; CI [0.1091-5.3179]; p = 0.7836). With a p > 0.05, this finding was not significant. The confidence

intervals were ranging from less than and greater than 1 showing no certainty of increase or decrease in hazard ratios.

Presence of comorbid cardiovascular diseases increased the hazard ratios as per this study (HR = 1.1482; CI [0.8925 – 1.4771]; p = 0.2824). The hazard ratio for an increase in patients' weight was 1.0099 (CI [0.9961 – 1.0239]; p = 0.160600). However, with a p > 0.05, this finding was not significant. The confidence intervals were ranging from less than and greater than 1, showing no certainty of increase or decrease in hazard ratios. Having secondary and tertiary education also was a significant predictor of DKD. Using primary education as the reference group, having secondary education was associated with a hazard ratio of 0.7503 (CI [0.5729 – 0.9826]; p = 0.0368), while tertiary education was linked to a hazard ratio of 0.5099 (CI [0.3555 - 0.7315]; p = 0.000254). Both levels of education significantly reduced the hazards of DKD in diabetic patients.

Patients who had spouses had increased hazard ratio (HR = 1.0716; CI [0.8019 - 1.4320]; p = 0.640078) taking those who had no spouses as point of reference. However, with a p > 0.05, this finding was not significant. The confidence intervals were ranging from less than and greater than 1 showing no certainty of increase or decrease in hazard ratios. In conjunction to that, Tobacco use was not a significant variable for DKD, with a hazard ratio of 0.7722 (CI [0.5340 - 1.1167]; p = 0.169578). In contrast, alcohol consumption had a positive effect on the hazard of DKD, with a hazard ratio of 1.5213 (CI [1.1528 - 2.0075]; p = 0.003027). Taking diabetic patients who did not use alcohol as reference, a patient who used alcohol increased their hazards of DKD by 52%. A family history of CKD significantly increased the hazard ratio by a factor of 12.1673 (p < 0.0001; CI [6.4701 – 22.8810]), using individuals with no family history as the reference group.

Rarely exercising increased the hazards of DKD by a factor 1.2637, taking frequently exercising patients as the point of reference. However, since p – value = 0.1703, it was not identified as a significant predictor of DKD. The confidence intervals were ranging from less than and greater than 1 showing no certainty of increase or decrease in hazard ratios. Being a retired diabetic patient (HR = 0.2791; CI [0.1819 - 0.4283]; p = < 0.0001)

or an unemployed patient (HR = 0.6534; CI [0.4280 - 0.9974]; p = 0.048599) reduced the hazards of DKD by 72.1% and 34.7%, respectively, taking the employed diabetic patients as the reference. Patients who were facing financial hardships had increased hazard ratio by a factor 5.1015; CI [0.8489 – 30.6596]; p = 0.044931 having taken those with no hardships as reference. The model yielded a likelihood ratio test result of 320.4, with a p-value of less than 2e-16.

Table 6: Summary results from multivariable Cox regression analysis (adjusted)

Variable	exp(coef)	lower 0.95	upper 0.95	p-value
Diabeticage	1.0486	1.0296	1.0680	3.92e-07
GenderMALE	1.3625	1.0506	1.7671	0.019697
HTNYES	0.7617	0.1091	5.3179	0.783643
CVDYES	1.1482	0.8925	1.4771	0.282382
Weight	1.0099	0.9961	1.0239	0.160600
EducationSECONDARY	0.7503	0.5729	0.9826	0.036848
EducationTERTIARY	0.5099	0.3555	0.7315	0.000254
SpouseYES	1.0716	0.8019	1.4320	0.640078
TobaccoYES	0.7722	0.5340	1.1167	0.169578
AlcoholYES	1.5213	1.1528	2.0075	0.003027
HistoryYES	12.1673	6.4701	22.8810	8.86e-15
ExerciseRARELY	1.2637	0.9044	1.7659	0.170302
EmploymentRETIRED	0.2791	0.1819	0.4283	5.22e-09
EmploymentUNEMPLOYED	0.6534	0.4280	0.9974	0.048599
FinancialhardshipYES	5.1015	0.8489	30.6596	0.044931

The current study concluded that given two diabetic patients, one male and the other female, the male patient has a higher risk of DKD by 36.25%. The higher risk of DKD in male diabetic patients compared to females can be attributed to genetic and epigenetic factors, hormonal differences, muscle structure, renal physiology, and lifestyle factors (Hsu & McCulloch, 2016; Kshirsagar & Dietrich, 2016; Gohda & Matsumoto, 2019; Mohlke & Boehnke, 2019; Mallea & Wang, 2021). Hecking *et al.,* (2014) suggested that gender differences in the prevalence of CKD might be attributed to several factors For instance, men generally have greater muscle mass than women, which can affect kidney function. Additionally, variations in hormone metabolism and glomerular structure between the sexes may contribute to the differences in CKD prevalence observed (Hecking *et al.,* 2014). While most studies suggest that both sexes are at risk for CKD related to diabetes, some research has found no significant relationship between sexes and

the risk or progression of CKD (Rossing *et al.,* 2004; Ricardo *et al.,* 2019). The findings of this study contradict those of Maric-Bilkan (2020), which indicated that women are at a higher risk for CKD than men. However, study by De Cosmo *et al.,* (2016) demonstrated that diabetic nephropathy is greater in men with type 2 diabetes and it therefore backs up the results in the current study.

The current results of this study indicated that hypertension was not a significant predictor of diabetic kidney disease (DKD). This finding contrasts with previous systematic reviews and meta-analyses that identified hypertension as a significant independent predictor of DKD (Noubiap *et al.,* 2015). For instance, Tekalign *et al*. (2023) showed that hypertensive patients are at six time's greater risk for DKD compared to those without hypertension. Additionally, another meta-analysis suggested that aggressive blood pressure-lowering strategies may be particularly beneficial for diabetic patients in preventing kidney failure (Wang *et al.,* 2019).

Possible reasons for the inconsistencies in this study may include the characteristics of the sample and the presence of confounding factors. For example, variations in age, sex, ethnicity, or socioeconomic status in the study populations may influence the prevalence and impact of hypertension on DKD. Studies with diverse populations might show different associations due to genetic, environmental, or lifestyle factors (Hsu & McCulloch, 2016). The health status and treatment history of participants can also affect findings. For instance, if the study population has different levels of hypertension control or comorbid conditions compared to other studies, this could alter the observed relationship between hypertension and DKD (Sanchez *et al*., 2015). Confounding factors may include medication adherence, lifestyle factors and other comorbid conditions.

The effectiveness of hypertension treatment can vary based on adherence to medication. Inadequate control of blood pressure due to poor adherence might weaken the observed association between hypertension and DKD (Krolewski & McClellan, 2013). Variables like diet, physical activity, and tobacco use can confound the relationship between hypertension and DKD. Lifestyle factors can influence both hypertension and kidney

health, potentially obscuring the direct effect of hypertension on DKD (Gohda & Matsumoto, 2019). Additional health conditions, like cardiovascular disease or obesity, can confound the relationship between hypertension and DKD. These comorbidities may interact with hypertension in complex ways, affecting kidney function independently of hypertension (Mallea & Wang, 2021).

The findings indicated that weight was not a significant predictor of diabetic kidney disease (DKD). This may be due to the understanding that weight alone does not directly influence the occurrence of DKD, as suggested by previous research. In contrast, the Body Mass Index (BMI)—calculated as weight relative to height squared - has been identified as a relevant factor in other studies. For example, Shiferaw *et al*. (2020) found that a BMI greater than 30 kg/m² was statistically linked to the incidence and progression of chronic kidney disease (CKD) in diabetic patients. An elevated BMI often indicates obesity, which is known to worsen conditions such as diabetes and hypertension, both of which are risk factors for DKD. Obesity can lead to inflammation and metabolic disturbances that accelerate kidney damage (Shiferaw *et al.,* 2020). It can also increase glomerular pressure and hyperfiltration, contributing to kidney injury. Moreover, excess body fat promotes systemic inflammation and insulin resistance, which further aggravates kidney function. Therefore, BMI serves as a more informative measure of these underlying processes than weight alone (Krolewski & McClellan, 2013). Consequently, relying solely on weight may not yield accurate predictions of DKD risk, as it does not adequately consider body composition and its related health effects.

The results of the current study indicate that diabetic patients who regularly consume alcohol have a greater risk of developing diabetic kidney disease (DKD) in comparison with those who do not. Chronic alcohol use can elevate blood pressure, which can damage the kidney's blood vessels, impairing their waste filtration ability (Gonzalez *et al*., 2017). Geng *et al.* (2020) found that simultaneous heavy drinking and diabetes significantly increase the hazards of end-stage kidney disease. Furthermore, insufficient moderate alcohol consumption has been linked to a higher likelihood of DKD (Roy *et al*., 2021). In contrast, Lee et al. (2021) found that alcohol intake was linked to a less

significant decline in estimated glomerular filtration rate (eGFR), suggesting a potential protective effect on kidney function in the general population. This finding contradicts the current study's conclusion that alcohol consumption elevates the risk of DKD. However, Lee *et al*. (2021) employed linear regression in their analysis, a common statistical approach with notable limitations, including sensitivity to outliers and the assumption of linearity. Outliers can significantly distort results, leading to misleading conclusions about how use of alcohol relates with kidney function (Rousseeuw & Leroy, 1987). Additionally, linear regression may not effectively capture complex interactions or non-linear relationships; if the impact of alcohol consumption on kidney function is non-linear, the method might not accurately represent this association (Fox, 2016).

In the current study, retirement or unemployment was linked to a decreased risk of diabetic kidney disease (DKD) among diabetic patients. This effect may be due to workplace-related stress, poor dietary habits, and long working hours. These findings align with those of Kim and Jang (2022), who reported that extended working hours are associated with chronic kidney disease (CKD) in full-time employees with diabetes, regardless of known risk factors. This relationship may stem from various factors related to stress, eating patterns, and work environments. Chronic stress has been shown to adversely affect health, including the progression of diabetes and its complications. Stress can lead to unhealthy coping strategies, such as poor dietary choices and inconsistent medication adherence, negatively impacting kidney health (Kim & Jang, 2022). Employees often have limited access to nutritious food and may opt for convenience foods high in sodium, sugar, and unhealthy fats, which can worsen metabolic control and kidney function (Kim & Jang, 2022). Conversely, retired individuals or those not employed may have more opportunities to engage in health-promoting activities, like regular exercise and preparing balanced meals, which can enhance overall health and potentially lower the risk of DKD (Kim & Jang, 2022). Reducing stress can lead to better blood glucose management and improved kidney function (Kim & Jang, 2022).

An analysis of variance was conducted for the parameters included in the model. The significant factors identified were age at diagnosis, gender, cardiovascular disease,

hypertension, weight, marital status, education level, family history of CKD, financial difficulties, employment status, alcohol consumption, and tobacco use (Table 7).The most statistically significant coefficients were hypertension ($p < 2.2e-16$) and history of chronic kidney disease ($p = 4.735e-13$) (Table 7). Unfortunately, the factor exercises in the study was determined to be statistically insignificant for the model ($p = 0.8858370$).

Table 7: Analysis of variance (ANOVA) for the Cox model

	Log likelihood	Chi Square	Df	Pr (>|Chi|)
NULL	-1514.8			
Diabetic age	-1509.9	9.8090	1	0.001737
Gender	-1506.7	6.5591	1	0.010435
HTN	-1469.0	75.3692	1	< 2.2e-16
CVD	-1457.8	22.2668	1	2.373e-06
Weight	-1440.1	35.3973	1	2.689e-09
Education	-1429.5	21.2547	2	2.424e-05
Spouse	-1426.4	6.2375	1	0.012507
Tobacco	-1409.5	33.7384	1	6.304e-09
Alcohol	-1401.5	15.9268	1	6.584e-05
History	-1375.4	52.3117	1	4.735e-13
Exercise	-1375.4	0.0206	1	0.885837
Employment	-1357.2	36.3334	2	1.289e-08
Financial hardship	-1354.6	5.2089	1	0.022471

Model selection was conducted using Akaike information criterion (AIC). Using a forward stepwise selection eight predictors were chosen for the model. In the model with the highest AIC of 2735.127, age at diabetes diagnosis, gender, employment status, history of CKD, financial hardships, education level, and alcohol use were identified as significant risk factors for DKD (Table 8). Predictors whose results revealed an increased risk of DKD included older age at diagnosis (HR=1.0534), male gender (HR = 1.3853), a family history of DKD (HR = 10.7401), financial hardships (HR = 4.0466), and alcohol use (HR = 1.6185). However, being unemployed (HR = 0.8048) or retired (HR = 0.2998), and having acquired secondary education (HR = 0.7309) or tertiary education (0.4880) showed a significant decrease in the overall model.

Table 8: Summary of the selected model using Akaike Information Criterion (AIC)

Variables	exp (coef)	lower 0.95	upper 0.95	Pr (> |z|)
Diabeticage	1.0534	1.0350	1.0721	6.93e-09
GenderMALE	1.3853	1.0870	1.7655	0.008440
EmploymentRETIRED	0.2998	0.1963	0.4579	2.48e-08
EmploymentUNEMPLOYED	0.8048	0.5432	1.1923	0.278845
HistoryYES	10.7401	6.3568	18.1460	< 2e-16
FinancialhardshipYES	4.0466	1.8795	8.7123	0.000353
EducationSECONDARY	0.7309	0.5613	0.9518	0.019982
EducationTERTIARY	0.4880	0.3447	0.6910	5.29e-05
AlcoholYES	1.6185	1.2348	2.1215	0.000488

The interpretation of the selected factors and their impacts on the risk of developing DKD was conducted using odds ratios (Table 8). The results indicated that for each additional year of age at the time of diabetes diagnosis, the risk of developing DKD increased by 5.34%. Males had a risk factor of 1.3853 for DKD compared to females, who served as the reference group. Additionally, being retired or unemployed reduced the risk of developing DKD by an average of 70% and 19%, respectively, compared to those who were employed, with employment serving as the reference level while holding other covariates constant (Table 8). A family history of CKD increased the risk by a factor of 10.74, using no family history as the reference group. When controlling for other factors and considering the absence of financial hardship as the reference, diabetic patients facing various financial difficulties had a risk of developing DKD that was 4.149 times higher than those without financial hardship. Patients with secondary education were 27% less likely to develop DKD, while those with tertiary education were 51.2% less likely, compared to individuals with only primary education or less. Finally, patients who consumed alcohol were 61% more likely to develop DKD compared to those who abstained.

In the selected model, a test for violations of the proportional hazards' assumption was conducted for the included covariates using the scaled Schoenfeld residuals test proposed by Grambsch and Therneau (1994). The results are presented in Table 9. The test for diabetic age (age at diabetes diagnosis) indicated marginal significance ($p = 0.0486$). Gender was found to be insignificant ($p = 0.4345$), thus not violating the proportional

hazards assumption. Employment was statistically significant ($p < 0.05$), indicating a violation of the assumption.

Other factors, including family history of CKD, financial hardship, and alcohol use, met the proportional hazards assumption of the Cox regression model, as they were not significant ($p > 0.05$) (see Table 9). However, the level of education was highly significant ($p = 0.0064$), which also represented a violation of the assumption. Additionally, the global test for the selected model supported the conclusion that the proportionality assumption was not upheld, as it was statistically significant ($p = 0.0073$). These violations suggest that the effects of certain covariates may change over time, potentially leading to misleading or erroneous results.

Table 9: Evaluation of the Proportional Hazards assumption for the selected model

Variable	Chisq	Df	P
Diabetic age	3.889	1	0.0486
Gender	0.619	1	0.4315
Employment	10.421	2	0.0055
History	0.110	1	0.7397
Financial hardship	0.202	1	0.6528
Education	10.095	2	0.0064
Alcohol	2.337	1	0.1263
GLOBAL	22.535	9	0.0073

The significant result of the global test suggests that, overall, the proportional hazards assumption for cox regression is violated for the model. This indicates that there are one or more covariates have effects on the hazard function that change over time, resulting in a lack of proportionality. The global test confirms that the violations are not limited to a few variables but are significant enough to affect the overall model's validity (Harrell, 2015). Diabetic age indicated a borderline violation of this assumption. This suggests that the effect of the duration of diabetes on DKD risk is not constant over time. For example, the impact of the length of time a person has had diabetes might vary with age, changes in health status, or the development of diabetes-related complications. Prior studies have also noted that the time with diabetes is not always a time-constant risk factor and can interact with other variables, such as glycaemic control and comorbid conditions, to

affect the progression of diabetic complications (Bianchi *et al*., 2021). Employment was significant variable, indicating that the effect of being employed or unemployed on DKD risk changes over time.

Employment status can be linked to varying levels of stress, lifestyle factors, and access to healthcare, all of which may fluctuate and interact differently over time, affecting DKD risk inconsistently (Kim & Jang, 2022). Education was significant variable, suggesting that its effect on DKD risk is also non-proportional over time. Education level may correlate with health literacy, socioeconomic status, and access to preventive healthcare, which can influence health behaviours and DKD risk variably across different periods (Fried *et al*., 2020). Higher education levels may initially provide a protective effect, but the impact could diminish or change as other factors, such as aging or comorbidities, come into play.

Gender, history, financial hardship, alcohol did not show significant p-values, indicating that their effects on DKD risk appear to remain proportional over time. For instance, gender differences for DKD may remain constant once adjusted for confounding factors like hormonal influences and baseline renal function (Hecking *et al.,* 2014).

The violation of the proportional hazards' assumption implies that the estimated hazard ratios for factors like diabetic age, employment, and education could be biased or misleading. This could lead to incorrect interpretations regarding the magnitude and direction of risk factors for DKD. For example, if the effect of employment changes over time but is assumed constant, the model may under- or overestimate the actual risk associated with employment status, potentially leading to flawed clinical or policy recommendations (Therneau & Grambsch, 2000). Violating this assumption can undermine the validity of findings, as the model may incorrectly represent the data. This can lead to erroneous conclusion about which factors are most critical in predicting DKD risk, affecting clinical decision-making and resource allocation for DKD prevention (Collett, 2015).

Adjustments for factor levels were made for employment and level of education so as to meet the proportional hazards assumption. Education levels (primary, secondary and tertiary) were adjusted to two levels; primary and below, secondary and above. Employment levels which consisted of employed, unemployed and retired, were adjusted to never employed (unemployed) and ever employed (employed and retired). The analysis was conducted similarly, including plots to check the log-rank tests for differences across various categories. The outcomes of the fitted Cox model are summarized in Table 10 and depicted in Figure 18.

Age at diabetes diagnosis had a positive effect on hazard of DKD (HR = 1.023; CI [1.01 – 1.04]; p = 0.001953). At any instance, a patient who was a year older at diagnosis had 2.3% more at risk of DKD than the younger patient. Being a male also significantly increased the hazards of DKD by 28% (HR = 1.2823; CI [1.01 – 1.63]; p = 0.0413). The findings showed that patients who were ever employed had a decreasing effect on the hazard ratio(HR) (HR=0.6354; CI [0.45 – 0.90]; p = 0.0106). Taking the unemployed as reference, having been in a form of employment reduced the hazards of DKD by 36.36% (Figure 18). Having a history of CKD produced the hazard ratio as well by a factor 6.9193; CI [4.22 - 11.36]; p < 0.00001. Here, having no family history of CKD was taken as the reference.

Taking no financial hardship as reference, presence of financial hardship among a diabetic patient increased their hazard by a factor 4.5244; CI [2.11 – 9.70]; p = 0.000104. Patients who were at least endowed with a secondary school education were at a less risk of DKD than those who had not received secondary school education by 40.68% (Figure 18).

Attaining secondary school education and above had a reducing effect on the hazard (HR= 0.5932; CI [0.46 – 0.76]; p = 0.0000326) taking primary school and below as reference. Use of alcohol increased the hazards of getting DKD among diabetic patients (HR = 0.1556; CI [1.19 – 2.04]; p = 0.001395. This was taking abstinence from alcohol use as reference.

The fitted model had an AIC= 2767.818. C-Index = 0.7822, se=0.01525 (Figure 18).

Table 10: Overview of the fitted Cox regression model.

Variables	Coeficient	exp (coef)	se (coef)	p- values
Diabeticage	0.022798	1.023060	0.007361	0.001953
GenderMALE	0.248688	1.282342	0.121860	0.041274
AdjEmploymentYES	-0.453474	0.635417	0.177372	0.010569
HistoryYES	1.934328	6.919393	0.252823	2.00e-14
FinancialhardshipYES	1.509493	4.524435	0.388868	0.000104
AdjEducationSECONDARY	-0.522186	0.593222	0.125687	3.26e-05
AlcoholYES	0.441933	1.555711	0.138286	0.001395

The hazard ratios for the cox model was illustrated as shown in Figure 18:

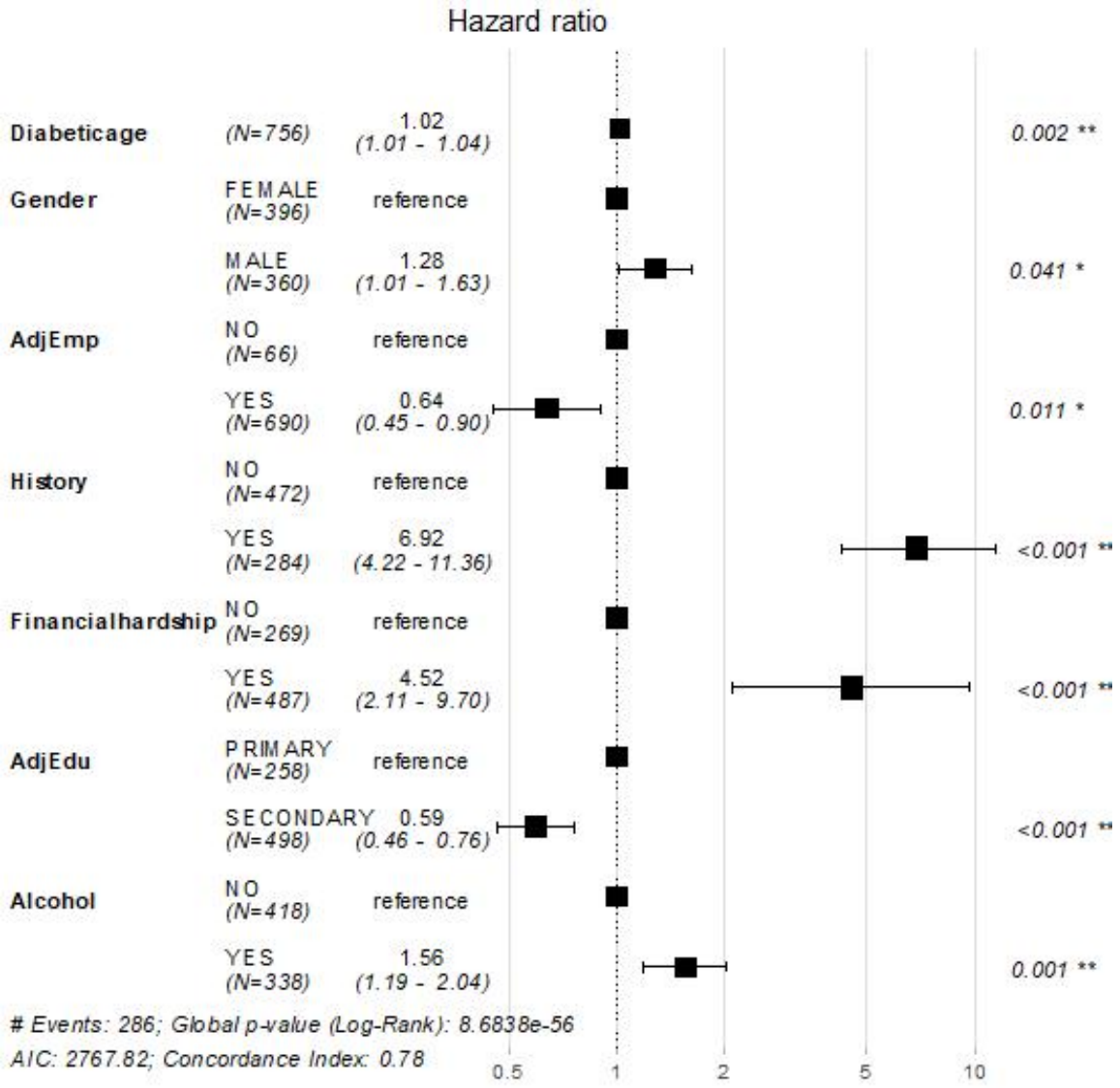

Figure 18: Effects of predictors of DKD and their confidence intervals

The fitted Cox proportional hazard model is expressed as follows:

ln(odds of hazard rate/ contracting DKD)= = 1.023060 Diabeticage + 1.282342 GenderMALE + 0.635417 AdjEmploymentYES + 6.919393 HistoryYES + 4.524435 FinancialhardshipYES + 0.593222 AdjEducationSECONDARY + 1.555711 AlcoholYES

The likelihood-ratio, Wald, and Score (Logrank) tests are asymptotically equivalent methods for assessing the omnibus null hypothesis that all the β coefficients are equal to zero (Table 11).

Table 11: Summary statistics for the fitted model

	Statistics	χ^2	Df	p-value
Fitted Model	Loglk	275.8	7	<2e-16
	Waldtest	155	7	<2e-16
	Logrank test	224.8	7	<2e-16

The global p-value of the fitted model (p = 0.268) indicated that it conforms to the proportional hazards' assumption (see Table 12). The predictors—diabetic age (p = 0.181), gender (p = 0.711), employment (p = 0.642), family history of CKD (p = 0.965), financial hardships (p = 0.458), education (p = 0.057), and alcohol use (p = 0.268)—were all statistically insignificant in the final model. Therefore, according to the scaled Schoenfeld's test, these predictors did not violate the proportional hazards assumption.

Table 12: Scaled Schoenfeld's Test for the fitted Cox model

	Chisq	Df	p
Diabeticage	1.78803	1	0.181
Gender	0.13755	1	0.711
Adj_Employment	0.21566	1	0.642
History	0.00191	1	0.965
FinancialHardship	0.55081	1	0.458
Adj_Education	3.96022	1	0.057
Alcohol	2.74770	1	0.097
GLOBAL	8.78816	7	0.268

Overall, the results from the scaled Schoenfeld's test indicate that the proportional hazards assumption is satisfied for the fitted Cox model, both globally and for individual predictors. This suggests that the model offers a reliable framework for understanding the factors that affect the risk of DKD over time in the studied population.

4.4.2 Support Vector Machine for Survival Analysis Model

Two survival SVM models (regression and hybrid) were developed using linear, additive, and radial basis function (RBF) kernels on the training dataset. Different values for the regularization parameter "gamma.mu" were tested. Following the fine-tuning of model parameters with 10-fold cross-validation, the models were trained. The C-index was employed to identify the top models. The mlr package was utilized for parameter tuning,

and the experiments were conducted on a high-performance computing platform. (Bischl *et al.,* 2016)

As anticipated, the choice of kernel function impacted the runtimes. The radial basis function (RBF) exhibited longer runtimes compared to the additive and linear kernel functions, which required approximately the same amount of time. This is due to the RBF needing an additional parameter to be tuned for optimal performance. Additionally, the hybrid survival SVM approach, which had the longest runtime among the methods, also reflects the influence of the number of parameters involved.

Table 13 displays the performance estimates for the survival support vector regression models. The estimates offer valuable insights into the effectiveness of various kernel functions and modeling approaches for predicting survival outcomes. The Concordance Index (C-index) is utilized to assess the model's discriminative ability, with a higher C-index signifying improved accuracy in predicting survival times. The regression approach using an additive kernel achieved the best performance recording a C-Index of 0.7753 (SD = 0.1). The hybrid approach with an additive kernel ranked second with a C – index of 0.7679 (SD = 0.09), followed closely by the hybrid approach with a linear kernel, which had ah C – index 0.7582 (SD = 0.1). The hybrid approach utilizing a radial basis function (RBF) kernel achieved a C-index of 0.7367 (SD = 0.04). The regression SVM models using a radial basis function kernel (C – index = 0.7317) and the linear kernel (C – index = 0.7300) were the lowest performing models in the current study. (Table 13)

Table 13: Performance estimates of survival support vector regression models

Type	Kernel	Concordance Index (C-index) (SD)
i) Regression	Linear	0.7300 (0.08)
	Additive	0.7753 (0.1)
	Radial basis function	0.7317 (0.06)
ii) Hybrid	Linear	0.7582 (0.1)
	Additive	0.7694 (0.09)
	Radial basis function	0.7367 (0.04)

In the regression model, while the linear kernel is recognized for its simplicity and computational efficiency, it may have difficulty effectively capturing complex, non-linear

relationships in survival data. The lower C-index indicates that the linear kernel is constrained in addressing more intricate survival patterns, which typically require more advanced modeling techniques for accurate outcome prediction (Grambsch & Therneau, 1994; Hastie *et al.,* 2009; Tong *et al.,* 2023). The additive kernel achieves a higher performance. This performance indicates that the additive kernel, which can model interactions between features while maintaining computational efficiency, offers superior predictive performance compared to the linear kernel. It is more effective at capturing non-linear relationships without greatly increasing the model's complexity (Schölkopf & Smola, 2002; Hastie *et al*., 2009; Tong *et al*., 2023). Radial Basis Function (RBF) Kernel showed a good performance. Although it is generally suitable for capturing non-linear relationships due to its flexibility, the slight improvement over the linear kernel indicates that in this case, the RBF kernel does not substantially outperform the simpler alternatives. The performance may be limited by the specific nature of the data or the additional computational burden required for parameter tuning (Bishop, 2006; Hastie *et al*., 2009; Ilemobayo *et al*., 2024).

In case of hybrid, the linear kernel attained an improvement over its performance in the regression model. This enhancement indicates that the hybrid approach, which integrates elements from multiple models or additional data components, enhances the linear kernel's performance. By leveraging complementary features or data interactions, the hybrid model likely enables the linear kernel to capture more complex relationships within the data, thereby improving its predictive accuracy (Hastie *et al*., 2009; Zhang & Zhao, 2017). In the hybrid model, the additive kernel achieved slightly lower than that observed in the regression model, which could be attributed to potential overfitting or the increased complexity introduced by the hybrid approach. Despite this slight decrease, the additive kernel continues to demonstrate strong performance, indicating its robustness and effectiveness across different modelling frameworks (Schölkopf & Smola, 2002; Hastie *et al*., 2009; Hem *et al*., 2021). Accordingly, for the hybrid model, RBF kernel achieved a slight improvement over its performance in the regression model. This suggests that the RBF kernel's performance benefits only marginally from the hybrid approach. The limited enhancement may be attributed to the RBF kernel's high sensitivity

to parameter tuning and the risk of overfitting, which can constrain its effectiveness even when integrated into a more complex modelling framework (Cristianini & Shawe-Taylor, 2000; Bishop, 2006; Bilal *et al*., 2024).

4.5 Performance of Cox Regression and Support Vector Machine Models

The runtimes for the survival support vector machine models were longer than those of the Cox regression model. To assess which method had the best predictive ability, the C-index was utilized, along with the AUC for the Cox model. The results indicated that the training set C-index for the Cox model was 0.7822, with an AUC of 0.721. For the validation set, the C-index was 0.770 and the AUC was 0.715. In comparison, the survival SVM regression approach using the additive kernel achieved a C-index of 0.7753, which was slightly better than that of the Cox model, indicating strong predictive capability. These findings are consistent with other research that suggests the accuracy of the Cox model, is somewhat lower than that of machine learning methods like SVM.

The comparison of performance between the Cox model and the SVM highlights significant differences in their predictive capabilities and computational efficiency. The Cox model demonstrated a training set C-index of 0.7822 and an AUC of 0.721, with slightly lower results for the validation set (C-index = 0.770, AUC = 0.715). In contrast, the survival SVM regression using the additive kernel achieved a slightly higher C-index of 0.7753, indicating marginally better predictive performance. However, it is worth noting that SVM models generally require longer processing times than the Cox model, despite their superior accuracy. This increased computational burden is often attributed to the complexity of SVMs, particularly when using kernel functions that necessitate additional parameter tuning (Cristianini & Shawe-Taylor, 2000; Hastie *et al*., 2009; Guido *et al*., 2024). The findings align with other studies that emphasize the benefits of machine learning methods, such as SVM, compared to semi-parametric models like the Cox model. Research has indicated that, although the Cox model offers important insights into survival data, its accuracy may be somewhat restricted when compared to advanced machine learning techniques, which are better equipped to identify complex

patterns and interactions within the data. (Van Houwelingen & Putter, 2012; Austin & Fine, 2017).

In summary, although the Cox model is a robust and interpretable tool for survival analysis, machine learning methods such as SVM can provide improved predictive performance, though they require more computational resources. The Cox Proportional Hazards Model, a traditional method in survival analysis, is appreciated for its reliability and ease of interpretation. It provides insights into the relationships between covariates and survival times, while maintaining a relatively straightforward computational approach. This model is based on the assumption of proportional hazards, making it suitable for scenarios where this assumption is valid. It provides valuable hazard ratios and allows for straightforward interpretation of the effects of predictors (Cox, 1972; Kalbfleisch & Prentice, 2002; Kuitunen *et al*., 2021).

However, machine learning techniques, like the SVM, have demonstrated superior predictive performance compared to traditional models like Cox regression (Xiao *et al*., 2022). Support Vector Machines can effectively capture intricate, non-linear relationships and interactions in the data thanks to their adaptable kernel functions. This allows SVMs to potentially identify patterns and dependencies that the Cox model might miss, especially in high-dimensional datasets. Despite their superior predictive ability, SVM come with increased computational demands. Training SVM models, particularly with non-linear kernels such as the RBF, can be time-consuming due to the need for solving complex optimization problems and performing extensive parameter tuning (Hastie *et al*., 2009).

CHAPTER FIVE
SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary

Diabetic kidney disease (DKD) is one of the primary causes of end-stage renal failure and poses a significant global public health challenge. As diabetes rates continue to increase globally, particularly in developing nations, the incidence of DKD is likely to rise as well, unless urgent enhancements are made in preventive clinical strategies. Diabetic Kidney Disease development and progression are influenced by various factors, making it crucial to identify relevant predictors and implement targeted preventive measures. Despite this, the understanding of DKD and the factors contributing to its onset remains limited.

Given the high prevalence of diabetes, accurately determining survival rates before the onset of DKD is crucial indicator for informing policymakers and physicians on improvement of treatment strategies. Meta-analysis and systematic reviews provide valuable pooled estimates, guiding the development of effective measures to enhance DKD survival rates and reduce diabetic mortality rates.

This study aimed to improve predictive modeling by integrating socio-economic factors to more accurately evaluate the survival rates of diabetic patients prior to developing diabetic kidney disease (DKD). The research employed both Cox models and Support Vector Machines (SVMs). A retrospective survey design was used to gather data from diabetes patients at Meru Teaching and Referral Hospital and Kerugoya Level 5 Hospital in Kenya. Information was collected on 13 variables, including age at diabetes diagnosis, gender, cardiovascular disease, hypertension, weight, marital status, educational attainment, family history of chronic kidney disease (CKD), financial difficulties, physical activity, employment status, alcohol consumption, and tobacco use.

R software was used for data analysis and model fitting due to its capability to handle and represent diverse types of the data. Various models within the software were employed to effectively analyse and interpret data. Log-rank tests were conducted on the categorical variables to obtain p-values and other relevant statistics, including expected values.

Kaplan-Meier curves were employed to graphically represent survival rates. Covariates with p-values below 0.05 were considered to have a significant effect on survival. A univariate Cox regression analysis was subsequently carried out to evaluate independent risk factors among the variables of interest. Additionally, multivariable Cox regression analysis was performed on all factors to assess their effects on hazard rates.

The findings indicated that approximately 40% of diabetic patients progressed to DKD, highlighting a significant prevalence of the condition among the cohort. Improving socio-economic welfare of diabetic patients could be a promising strategy for enhancing diabetes treatment and delaying DKD onset. However, challenges such as relevant education, unemployment, inflation and high cost of living have impeded the implementation of such measures.

The study's findings showed considerable diversity in patient characteristics. The duration of diabetes among participants varied from 1 to 31 years, with a median of 12 years. This indicates that the group includes both newly diagnosed patients and those with a long-standing history of the disease, highlighting a wide range of disease progression and management experiences. Patients' ages ranged from 15 to 76 years, with a median age of 41, suggesting that a significant number are middle-aged. Additionally, weights varied from 56 to 116 kg, with a median of 85 kg, indicating that many patients may be classified as obese.

The study found that a considerable percentage of diabetic patients had hypertension (64.3%) and comorbid cardiovascular diseases (27.2%). Educational attainment was diverse, with 33.3% having completed primary education, 36.8% secondary education, and 29.9% tertiary education. About 51% of the patients were without spouses due to various reasons. Employment status varied, with 56.9% employed, 8.7% unemployed, and 34.4% retired. Financial hardship was reported by 56.1% of participants. Additionally, 65.3% had never used tobacco, and 65.9% had no history of alcohol use.

In the model, the significant variables identified were age at diagnosis, gender, level of education, family history of CKD, financial hardship, employment, alcohol use and tobacco use. However, cardiovascular disease, hypertension, weight, marital status and physical exercises did not show statistical significance. The final model, selected based on AIC included seven predictors, namely diabetic age, gender, employment, family history of CKD, financial hardship, level of education and alcohol use.

A survival support vector machine (SVM) model was trained and tested for comparison. The predictive accuracy of the SVM regression model was assessed against the Cox regression model using the C-index as the performance metric. The SVM model achieved a marginally higher C-index of 0.7753, compared to 0.770 for the Cox model. This suggests that the SVM model has a slight edge in predictive accuracy over the Cox model, indicating that it may be more effective in identifying patterns and predicting survival outcomes in this context. However, the difference is minimal, so while the SVM model shows potential for improved prediction, the Cox model still remains a strong competitor in survival analysis.

The study suggests that the overall socio-economic circumstances of diabetic patients' significantly affect their wellbeing. Moreover, managing diabetes is inherently challenging, requiring patients to maintain strict discipline, frequently seek consultations doctors and follow up necessary treatment protocols. Modeling can be essential for comprehending and managing health outcomes in diabetic patients, particularly regarding the onset and progression of diabetic kidney disease (DKD). Modelling provides a robust framework for integrating clinical, demographic, and socio-economic data to predict outcomes, tailor interventions and optimize resource use. Ultimately, this can enhance the quality of care for diabetic patients at risk of developing diabetic kidney disease (DKD).

5.2 Conclusion

The research showed that diabetic kidney disease (DKD) is quite common among diabetes patients in Kenya, affecting approximately 38% of them. This study identified

several significant predictors of DKD including age at diagnosis, gender, family history of CKD, use of alcohol, financial hardship, employment status and level of education.

The Cox regression model identified various predictors linked to a higher risk of diabetic kidney disease (DKD), such as age at diagnosis, male gender, a family history of CKD, use of alcohol and financial hardship. In contrast, being employed and having completed secondary education or higher were significantly associated with a lower risk of DKD in individuals with diabetes. These results deepen the understanding of DKD risk factors in diabetic patients and support the creation of targeted prevention strategies. Therefore, it is essential to develop tailored interventions that address the specific contexts of diabetes patients to reduce both the prevalence and risk factors of DKD.

The study also confirmed that the survival support vector regression model using an additive kernel marginally outperforms the Cox regression model in predictive ability. However, limitations were encountered, including inconsistent and inaccurate data reported by DKD and diabetic patients as well as incomplete hospital records. These issues led to the exclusion of several observations due to due to inconsistencies, which may have affected the representation of the actual situation.

5.3 Recommendations

Based on the results of this study, the following recommendations are proposed:

i. Use Support Vector Machine for survival analysis of censored data as it has a higher predictive accuracy.
ii. Implement more advanced recording of patients file to prevent the paucity of medical research data.
iii. Improving socio-economic conditions, such as providing financial support, creating of employment opportunities and ensuring access to quality education, may assist reduce progression of various chronic illnesses in diabetic patients.
iv. Use the developed models in predicting diabetic patients' risk of DKD and in decision making in order to control kidney disease occurrence.

5.4 Suggestion for Further Studies

Based on the findings of this study, the following areas are suggested for further research

i. Identifying the most effective prediction methods for small samples.

ii. Exploring alternative performance metrics beyond the C -index for evaluating the performance of a Support Vector Machine with censored survival data.

REFERENCES

Abd, E. S., Bolignano, D., D'Arrigo, G., Dounousi, E., Tripepi, G., & Zoccali, C. (2018). Prevalence and Burden of Chronic Kidney Disease among the General Population and High-Risk Groups in Africa: A systematic review. *BMJ Open*, *8*(1), e015069.

Afkarian, M., Sachs, M. C., Kestenbaum, B., Hirsch, I. B., Tuttle, K. R., Himmelfarb, J., & de Boer, I. H. (2013). Kidney Disease and Increased Mortality Risk in Type 2 Diabetes. *Journal of the American Society of Nephrology*, *24*(2), 302–308.

Aghaabbasi, M., & Chalermpong, S. (2023). Machine learning techniques for evaluating the nonlinear link between built-environment characteristics and travel behaviors: A systematic review. *Travel Behaviour and Society*, *33*, 100640.

Allen, A., Iqbal, Z., Green-Saxena, A., Hurtado, M., Hoffman, J., Mao, Q., & Das, R. (2022). Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ Open Diabetes Research & Care*, *10*(1), e002560.

Amaya-Tejera, N., Gamarra, M., Vélez, J.I., & Zurek, E. (2024). A distance-based kernel for classification via support vector machines. *Frontiers in Artificial Intelligence*, *7*.

American Diabetes Association. (2010). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, *33*(Supplement_1), S62–S69.

Anderson, J. W., Kendall, C. W. C., & Jenkins, D. J. A. (2005). Importance of Weight Management in Type 2 Diabetes: Review with Meta-analysis of Clinical Studies. *Journal of the American College of Nutrition*, 24(5), 431-439.

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, *6*, 100094.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioural Research*, 46(3), 399-424.

Austin, P. C., & Fine, J. P. (2017). Practical Recommendations for Reporting Fine-Gray Model Results for Competing Risks. *Statistics in Medicine*, 36(27), 4391-4400.

Bell, M. L., Whitehead, A. L., & Julious, S. A. (2018). Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. *Clinical Epidemiology*, *Volume 10*, 153–157.

Berkowitz, S. A., Meigs, J. B., DeWalt, D., Seligman, H. K., Barnard, L. S., Bright, O.-J. M., Schow, M., Atlas, S. J., & Wexler, D. J. (2015). Material Need Insecurities,

Control of Diabetes Mellitus, and Use of Health Care Resources. *JAMA Internal Medicine*, *175*(2), 257.

Bianchi, S., Bigazzi, R., & Caiazza, A. (2021). "Effect of Duration of Diabetes on Diabetic Nephropathy: A Population-Based Study." *Diabetes Research and Clinical Practice*.

Bihan, H., Laurent, S., Sass, C., Nguyen, G., Huot, C., Moulin, J. J., Guegen, R., Le Toumelin, P., Le Clesiau, H., La Rosa, E., Reach, G., & Cohen, R. (2005). Association among Individual Deprivation, Glycemic Control, and Diabetes Complications: The EPICES score. *Diabetes Care*, *28*(11), 2680–2685.

Bihan, H., Ramentol, M., Fysekidis, M., Auclair, C., Gerbaud, L., Desbiez, F., Peyrol, F., Thieblot, P., Cohen, R., & Tauveron, I. (2012). Screening for deprivation using the EPICES score: a tool for detecting patients at high risk of diabetic complications and poor quality of life. *Diabetes & metabolism*, *38*(1), 82–85.

Bikbov, B., Purcell, C. A., Levey, A. S., Smith, M., Abdoli, A., Abebe, M., Adebayo, O. M., Afarideh, M., Agarwal, S. K., Agudelo-Botero, M., Ahmadian, E., Al-Aly, Z., Alipour, V., Almasi-Hashiani, A., Al-Raddadi, R. M., Alvis-Guzman, N., Amini, S., Andrei, T., Andrei, C. L., & Andualem, Z. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *395*(10225), 709–733.

Bilal, A., Imran, A., Baig, T. I., Liu, X., Abouel Nasr, E., & Long, H. (2024). Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. *Scientific Reports*, *14*(1).

Bishop, C. (2007) Pattern Recognition and Machine Learning. *Springer-Verlag,* 2nd edition.

Bland, J. M., & Altman, D. G. (2004). The logrank test. *BMJ*, *328*(7447), 1073.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, *89*(3), 431–436.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Brown, A. F., Ettner, S. L., Piette, J., Weinberger, M., Gregg, E., Shapiro, M. F., Karter, A. J., Safford, M., Waitzfelder, B., Prata, P. A., & Beckles, G. L. (2004). Socioeconomic position and health among persons with diabetes mellitus: a conceptual framework and review of the literature. *Epidemiologic reviews*, *26*, 63–77.

Buffet, L., & Ricchetti, C. (2012, June 20). Chronic Kidney Disease and Hypertension: A DestructiveCombination.*Uspharmacist.com*.

Centers for Disease Control and Prevention (CDC). (2004). Prevalence of overweight and obesity among adults with diagnosed diabetes--United States, 1988-1994 and 1999-2002. *MMWR. Morbidity and mortality weekly report*, *53*(45), 1066-1068.

Cervantes, J., Li, X., Yu, W., & Li, K. (2008). Support vector machine classification for large data sets via minimum enclosing ball clustering. *Neurocomputing*, *71*(4–6), 611–619.

Chagnac, A., Weinstein, T., Herman, M., Hirsh, J., Gafter, U., & Ori, Y. (2003). The effects of weight loss on renal function in patients with severe obesity. *Journal of the American Society of Nephrology: JASN*, *14*(6), 1480–1486.

Chen, I-Ju., Chuang, Y.-H., Hsu, L.-T., & Chen, J.-Y. (2020). *Association between Marital Status and Chronic Kidney Disease among Middle-Aged and Elderly Taiwanese: A Community-Based, Cross-Sectional Study*. *14*(3), 174–178.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. I. D. F. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, *138*, 271-281.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, *110*, 12-22.

Cohen, S. D., Sharma, T., Acquaviva, K., Peterson, R. A., Patel, S. S., & Kimmel, P. L. (2007). Social support and chronic kidney disease: an update. *Advances in chronic kidney disease*, *14*(4), 335–344.

Colberg, S. R., Sigal, R. J., Yardley, J. E., et al. (2016). Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association. *Diabetes Care, 39*(11), 2065-2079.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC Press.

Colosimo, E., Ferreira, F. V., Oliveira, M., & Sousa, C. (2002). Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation*, *72*(4), 299-308.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Crook, O.M., Chung, C., & Deane, C.M. (2022). Challenges and opportunities for Bayesian statistics in Proteomics. *Journal of Proteome Research*, *21*(4), 849-864.

Cutler, D. M., & Lleras-Muney, A. (2010). Understanding differences in health behaviours by education. Journal of Health Economics, 29(1), 1-28.

Cutler, S. J., & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of chronic diseases*, *8*(6), 699-712.

De Graft-Johnson, J., & Schneider, K. (2020). Employment status and health outcomes: An examination of chronic disease management and survival. Social Science & Medicine, 258, 113177.

Distiller, L. A. (2014). Why do some patients with type 1 diabetes live so long? *World journal of diabetes*, *5*(3), 282–287.

Dovgan, E., Gradišek, A., Luštrek, M., Uddin, M., Nursetyo, A. A., Annavarajula, S. K., Li, Y.-C., & Syed-Abdul, S. (2020). Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. *PLOS ONE*, *15*(6), e0233976.

Dunkler, D., Gao, P., Lee, S. F., Heinze, G., Clase, C. M., Tobe, S., & Oberbauer, R. (2015). Risk prediction for early CKD in type 2 diabetes. *Clinical Journal of the American Society of Nephrology*, *10*(8), 1371-1379.

Duru, O. K., Middleton, T., Tewari, M. K., & Norris, K. (2018). The Landscape of Diabetic Kidney Disease in the United States. *Current Diabetes Reports*, *18*(3).

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., Deprosto, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019) A guide to deep learning in healthcare. Nat Med 25, 24–29 (2019).

Federation, I. D. (2019). Idf diabetes atlas. 2013. *International Diabetes Federation.* ehp. v119. i03.

Fenta, E. T., Eshetu, H. B., Kebede, N., Bogale, E. K., Zewdie, A., Kassie, T. D., Anagaw, T. F., Mazengia, E. M., & Gelaw, S. S. (2023). Prevalence and predictors of chronic kidney disease among type 2 diabetic patients' worldwide, systematic review and meta-analysis. *Diabetology & metabolic syndrome*, *15*(1), 245.

Fiore, M. C., & Jaén, C. R. (2020). Treating Tobacco Use and Dependence: 2008 Update. *Clinical Practice Guidelines*. *Journal of the American Medical Association, 303*(5), 639-640.

Fouodo, C., König, I., Weihs, C., Ziegler, A., & Wright, M. (2018). Support vector machines for survival analysis with R. *The R Journal*, *10*(1), 412.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Fried, L., Storer, T., & Glidden, D. (2020). "Education and Health Literacy in CKD Risk: A Meta-Analysis." *Journal of Nephrology*.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189-1232.

Funakoshi, M., Azami, Y., Matsumoto, H., Ikota, A., Ito, K., Okimoto, H., Shimizu, N., Tsujimura, F., Fukuda, H., Miyagi, C., Osawa, S., Osawa, R., & Miura, J. (2017). Socioeconomic Status and Type 2 Diabetes Complications among Young Adult Patients in Japan. *PloS one*, *12*(4), e0176087.

Gachoki, P., Muraya, M. M., Njoroge G. G. (2022b) Features Selection in Statistical Classification of High Dimensional Image Derived Maize (*Zea Mays* L.) Phenomic Data. *American Journal of Applied Mathematics and Statistics*, 2, 44-51

Gachoki, P., Muraya, M. M., Njoroge, G. G. (2022a) Modelling Plant Growth Based on Gompertz, Logistic Curve, Extreme Gradient Boosting and Light Gradient Boosting Models Using High Dimensional Image Derived Maize (*Zea mays* L.) Phenomic Data. *American Journal of Applied Mathematics and Statistics*, 10(2):52-64

Gajewska, K. A., Bennett, K., Biesma, R., & Sreenan, S. (2020). Low uptake of continuous subcutaneous insulin infusion therapy in people with type 1 diabetes in Ireland: a retrospective cross-sectional study. *BMC Endocrine Disorders*, *20*(1).

GBD 2021 Diabetes Collaborators (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet (London, England)*, *402*(10397), 203–234.

Geletu, A. H., Teferra, A. S., Sisay, M. M., & Teshome, D. F. (2018). Incidence and predictors of chronic kidney diseases among type 2 diabetes mellitus patients at St. Paul's Hospital, Addis Ababa, Ethiopia. *BMC research notes*, *11*, 1-6.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

Geng, T. T., Jafar, T. H., Yuan, J. M., & Koh, W. P. (2020). The impact of diabetes on the association between alcohol intake and the risk of end-stage kidney disease in the Singapore Chinese Health Study. Journal of diabetes, 12(8), 583–593.

Geogrousopoulou, E. N., Pitsavos, C., Yannakoulia, M., & Panagiotakos, D. B. (2015). Comparisons between Survival Models in Predicting Cardiovascular Disease Events: Application in the ATTICA Study (2002-2012). *Journal of Statistics Applications & Probability*, *4*(2), 203–210.

Gibbons, M. E., & Kelsey, J. L. (2022). The Impact of Financial Hardship on Chronic Disease Management and Health Outcomes. *Health Economics Review, 12*(1), 45.

Gohda, T., & Matsumoto, M. (2019). "Sex differences in diabetic kidney disease." *Clinical and Experimental Nephrology*, 23(4), 434-441.

Gonzalez, H., & Adams, R. D. (2017). "Alcohol Consumption and Hypertension: A Review of the Literature." Hypertension Research, 40(1), 1-10.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515-526.

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An overview on the advancements of support vector machine models in healthcare applications: A review. *Information*, *15*(4), 235.

Gündoğdu, Y., & Anaforoğlu, İ. (2022). Effects of Smoking on Diabetic Nephropathy. *Frontiers in clinical diabetes and healthcare*, *3*, 826383.

Harrell, F. E. (2015). Regression Modeling Strategies. In *Springer Series in Statistics*. Springer International Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hecking, M., Bieber, B. A., Ethier, J., Kautzky-Willer, A., Sunder-Plassmann, G., Säemann, M. D., & Port, F. K. (2014). Sex-specific differences in hemodialysis prevalence and practices and the male-to-female mortality rate: the Dialysis Outcomes and Practice Patterns Study (DOPPS). *PLoS medicine*, *11*(10), e1001750.

Hecking, M., Bieber, B. A., Ethier, J., Kautzky-Willer, A., Sunder-Plassmann. (2014). "Differences in CKD Prevalence by Gender: A Hormonal and Structural Perspective." *Kidney International*.

Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G., & Riebler, A. (2021). Robust modelling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics*, *217*(3).

Herbrich, R., Graepel, T., & Obermayer, K. (1999). *Support vector learning for ordinal regression*. IEEE Xplore.

Hoogeveen, E. K. (2022). The epidemiology of diabetic kidney disease. *Kidney and Dialysis*, *2*(3), 433-442.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Hsu, C. Y., & McCulloch, C. E. (2016). "Sex and gender differences in diabetes and kidney disease." *Current Diabetes Reports*, 16(8), 54.

Hu, C., & Steingrimsson, J. A. (2017). Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. *Journal of Biopharmaceutical Statistics*, *28* (2), 333–349.

Ilemobayo, J.A., Durodola, O., Alade, O., J Awotunde, O., T Olanrewaju, A., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., E Odezuligbo, I., & E Edu, O. (2024). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. Journal of Engineering Research and Reports, 26(6), 388–395.

Ishwaran, H., Kogalur, U. B., Chen, X., & Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, *4*(1), 115–132.

Jager, K. J., Kovesdy, C., Langham, R., Rosenberg, M., Jha, V., & Zoccali, C. (2019). A single number for advocacy and communication—worldwide more than 850 million individuals have kidney diseases. *Nephrology Dialysis Transplantation*, *34*(11), 1803-1805.

Jiang, W., Wang, J., Shen, X., Lu, W., Wang, Y., Li, W. & Chang, B. (2020). Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. *Diabetes Care*, *43*(4), 925-933.

Johnson, L. L., & Shih, J. H. (2007). An introduction to survival analysis. In *Principles and practice of clinical research* (pp. 273-282). Academic Press.

Kairu, B. N. (2015). *Prevalence of chronic kidney disease among ambulatory Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome patients on antiretroviral therapy at the Kenyatta National Hospital*. Erepository.uonbi.ac.ke. http://erepository.uonbi.ac.ke/handle/11295/95105

Kalantar-Zadeh, K., Jafar, T. H., Nitsch, D., Neuen, B. L., & Perkovic, V. (2021). Chronic kidney disease. *Lancet (London, England)*, *398*(10302), 786–802.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Ke, C., Kim, S. J., Shah, B. R., Bierman, A. S., Lipscombe, L. L., Feig, D. S., & Booth, G. L. (2019). Impact of Socioeconomic Status on Incidence of End-Stage Renal Disease and Mortality after Dialysis in Adults with Diabetes. *Canadian journal of diabetes*, *43*(7), 483–489.e4.

Kim, K.D., & Jang, S. (2022). Association between long working hours and chronic kidney disease according to diabetic status. *Journal of Occupational & Environmental Medicine*, *64*(3), 190-196.

Kim, Y., Kim, W., Kim, J., Moon, J.Y., Park,S., Park, C.W., Park, H.S., Song, S.H., Yoo, T., Lee, S., Lee, E.Y., Lee, J., Jin, K., Cha, D. R., Cha, J. J., & Han, S. Y. (2022). Blood pressure control in patients with diabetic kidney disease. *Electrolytes & Blood Pressure*, *20*(2), 39.

Kleinbaum, D. G., & Klein, M. (2012). Survival Analysis. In *Statistics for Biology and Health*. Springer New York.

Kovesdy, C. P., & Liew, A. (2018). The Role of Genetic Factors in Chronic Kidney Disease Progression. *Kidney International Supplements, 8*(1), 22-28.

Krishnamurthy, S., Kapeleshh, K., Dovgan, E., Luštrek, M., Gradišek Piletič, B., Srinivasan, K., Li, Y. J., Gradišek, A., & Syed-Abdul, S. (2021). Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan. *Healthcare (Basel, Switzerland)*, *9*(5), 546.

Krolewski, A. S., & McClellan, W. (2013). "Sex differences in the risk of diabetic kidney disease." Diabetes Care, 36(4), 863-872.

Kshirsagar, V., & Dietrich, R. (2016). "Impact of muscle mass on kidney function in diabetic patients." *Journal of Diabetes and its Complications*, 30(7), 1244-1250.

Kuitunen, I., Ponkilainen, V. T., Uimonen, M. M., Eskelinen, A., & Reito, A. (2021). Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: Methodological perspectives and review. *BMC Musculoskeletal Disorders*, *22*(1).

Kumar, M., Dev, S., Khalid, M. U., Siddenthi, S. M., Noman, M., John, C., Akubuiro, C., Haider, A., Rani, R., Kashif, M., Varrassi, G., Khatri, M., Kumar, S., &

Mohamad, T. (2023). The bidirectional link between diabetes and kidney disease: Mechanisms and management. *Cureus*.

Lee, E.T., & Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, ISBN 978-0-471-45854-8. [p412]

Lee, Y.-J., Cho, S., & Kim, S. R. (2021). Effect of alcohol consumption on kidney function: population-based cohort study. *Scientific Reports*, *11*(1).

Lin, H. C., Peng, C. H., Chiou, J. Y., & Huang, C. N. (2014). Physical activity is associated with decreased incidence of chronic kidney disease in type 2 diabetes patients: a retrospective cohort study in Taiwan. *Primary care diabetes*, *8*(4), 315-321.

Lin, J., Zhang, D., & Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics*, *62*(3), 803-812.

Liu, L., Xia, R., Song, X., Zhang, B., He, W., Zhou, X., Li, S., & Yuan, G. (2021). *Association between the triglyceride–glucose index and diabetic nephropathy in patients with type 2 diabetes: A cross-sectional study*. *12*(4), 557–565.

Liu, M., Li, X. C., Lu, L., Cao, Y., Sun, R. R., Chen, S., & Zhang, P. Y. (2014). Cardiovascular disease and its relationship with chronic kidney disease. *European review for medical and pharmacological sciences*, *18*(19), 2918–2926.

Mallea, J., & Wang, J. (2021). "Gender differences in glomerular filtration rate and diabetic kidney disease." *Kidney International Reports*, 6(4), 965-973.

Maric-Bilkan, C. (2020, March). Sex differences in diabetic kidney disease. In *Mayo Clinic Proceedings* (Vol. 95, No. 3, pp. 587-599). Elsevier.

Mills, K. T., Xu, Y., Zhang, W., Bundy, J. D., Chen, C. S., Kelly, T. N., & He, J. (2015). A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010. *Kidney international*, *88*(5), 950-957.

Mohammedi, K., Chalmers, J., Herrington, W., Li, Q., Mancia, G., Marre, M., Poulter, N., Rodgers, A., Williams, B., Perkovic, V., Coresh, J., & Woodward, M. (2018). Associations between body mass index and the risk of renal events in patients with type 2 diabetes. *Nutrition & Diabetes*, *8*(1).

Mohlke, K. L., & Boehnke, M. (2019). "Genetic variants associated with diabetic kidney disease: a review." *Journal of Nephrology*, 32(2), 163-170.

More, K. S., & Wolkersdorfer, C. (2023), Application of machine learning algorithms for nonlinear system forecasting through analytics — A case study with mining influenced water data. Water Resources and Industry, 29.

Mu, X., Wu, A., Hu, H., Zhou, H., & Yang, M. (2023). Prediction of Diabetic Kidney Disease in Newly Diagnosed Type 2 Diabetes Mellitus. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *Volume 16*, 2061–2075.

Mule, G., Castiglia, A., Cusumano, C., Scaduto, E., Geraci, G., Altieri, D., & Cottone, S. (2017). Subclinical kidney damage in hypertensive patients: a renal window opened on the cardiovascular system. Focus on microalbuminuria. *Hypertension: from basic research to clinical practice*, 279-306

Muntner, P., & Judd, S. E. (2017). Strategies for managing hypertension in patients with chronic diseases: Evidence from recent studies. *American Journal of Medicine, 130*(5), 493-501.

Mwaura, J. (2024). Diabetes in Kenya: latest updates in 2024. *labtestzote.com*.

Nakashima, A., Kato, K., Ohkido, I., & Yokoo, T. (2021). Role and Treatment of Insulin Resistance in Patients with Chronic Kidney Disease: A Review. *Nutrients*, *13*(12), 4349.

Nayak S., Amin A., Reghunath S.R:, Thunga G, Dinesh A.U., Shivashankara K.N., Attur R.P., Acharya L.D., (2024) Development of a machine learning-based model for the prediction and progression of diabetic kidney disease: A single centred retrospective study, International Journal of Medical Informatics, Volume 190,

O'Neill, A. (2024). *Unemployment rate in Kenya 2023*. Statista; Statista.

Obare. D. M., & Muraya. M.M., (2018) Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. *American Journal of Applied Mathematics and Statistics*. 6 (6): 266 - 271.

Orth, S. R. (2000). Smoking–a renal risk factor. *Nephron*, *86*(1), 12-26.

Pacilli, A., Viazzi, F., Fioretto, P., Giorda, C., Ceriello, A., Genovese, S., & AMD-Annals Study Group. (2017). Epidemiology of diabetic kidney disease in adult patients with type 1 diabetes in Italy: The AMD-Annals initiative. *Diabetes/metabolism research and reviews*, *33*(4), e2873.

Pagano, M., Gauvreau, K., & Mattie, H. (2022). *Principles of biostatistics* (3rd ed.). Chapman and Hall/CRC.

Parizadeh, D., Rahimian, N., Akbarpour, S., Azizi, F., & Hadaegh, F. (2019). Sex-specific clinical outcomes of impaired glucose status: A long follow-up from the Tehran Lipid and Glucose Study. *European journal of preventive cardiology*, *26*(10), 1080–1091.

Parving, H. H., Lewis, J. B., Ravid, M., Remuzzi, G., & Hunsicker, L. G. (2006). Prevalence and risk factors for microalbuminuria in a referred cohort of type II diabetic patients: a global perspective. *Kidney international*, *69*(11), 2057-2063.

Phillips, K., Hazlehurst, J. M., Sheppard, C., Bellary, S., Hanif, W., Karamat, M. A., Crowe, F. L., Stone, A., Thomas, G. N., Peracha, J., Fenton, A., Sainsbury, C., Nirantharakumar, K., & Dasgupta, I. (2024). Inequalities in the management of diabetic kidney disease in UK primary care: A cross-sectional analysis of a large primary care database. *Diabetic medicine: a journal of the British Diabetic Association*, *41*(1), e15153.

Pippitt, K., Li, M., & Gurgle, H. E. (2016). Diabetes Mellitus: Screening and Diagnosis. *American family physician*, *93*(2), 103–109.

Pöhlmann, J., Bergenheim, K., Garcia Sanchez, J. J., Rao, N., Briggs, A., & Pollock, R. F. (2022). Modeling Chronic Kidney Disease in Type 2 Diabetes Mellitus: A Systematic Literature Review of Models, Data Sources, and Derivation Cohorts. *Diabetes therapy: research, treatment and education of diabetes and related disorders*, *13*(4), 651–677.

Rabkin, R. (2003). Diabetic nephropathy. *Clinical cornerstone*, *5*(2), 1–11.

Rhee, C. M., & Kovesdy, C. P. (2015). Spotlight on CKD deaths—increasing mortality worldwide. *Nature Reviews Nephrology*, *11*(4), 199-200.

Ricardo, A. C., Yang, W., Sha, D., Appel, L. J., Chen, J., Krousel-Wood, M., ... & CRIC Investigators. (2019). Sex-related disparities in CKD progression. *Journal of the American Society of Nephrology*, *30*(1), 137-146.

Rossing, K., Christensen, P. K., Hovind, P., Tarnow, L., Rossing, P., & Parving, H. H. (2004). Progression of nephropathy in type 2 diabetic patients. *Kidney international*, *66*(4), 1596-1605.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

Roy, S., Schweiker-Kahn, O., Jafry, B., Masel-Miller, R., Raju, R. S., O'Neill, L. M. O., Correia, C. R., Trivedi, A., Johnson, C., Pilot, C., Saddemi, J., Memon, A., Chen, A., McHugh, S. P., Patel, S., Daroshefski, N. M., Nguyen, T., Wissler, W., Sharma, E., & Hunter, K. (2021). Risk Factors and Comorbidities Associated with Diabetic Kidney Disease. *Journal of primary care & community health*, *12*, 21501327211048556.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.

Sabanayagam, C., He, F., Nusinovici, S., Li, J., Lim, C., Tan, G., & Cheng, C. Y. (2023). Prediction of diabetic kidney disease risk using machine learning models: a population-based cohort study of Asian adults. *ELife*, *12*, e81878.

Saeed, M., Stene, L. C., Reisæter, A. V., Jenssen, T. G., Tell, G. S., Tapia, G., Joner, G., & Skrivarhaug, T. (2022). End-stage renal disease: incidence and prediction by coronary heart disease, and educational level. Follow-up from diagnosis of childhood-onset type 1 diabetes throughout Norway 1973-2017. *Annals of epidemiology*, *76*, 181–187.

Saegusa, T., Di, C., & Chen, Y. Q. (2014). Hypothesis testing for an extended cox model with time-varying coefficients. *Biometrics*, *70*(3), 619-628.

Saiki, A., Nagayama, D., Ohhira, M., Endoh, K., Ohtsuka, M., Koide, N., Oyama, T., Miyashita, Y., & Shirai, K. (2005). Effect of weight loss using formula diet on renal function in obese patients with diabetic nephropathy. *International journal of obesity (2005)*, *29*(9), 1115–1120.

Sanchez, O. A., Jacobs, D. R., Bahrami, H., Peralta, C. A., Daniels, L. B., Lima, J. A., Maisel, A., & Duprez, D. A. (2015). Increasing aminoterminal-pro-B-type natriuretic peptide precedes the development of arterial hypertension. *Journal of Hypertension*, *33*(5), 966–974.

Saya, M. (2023). KU hospital doubles daily dialysis patients after acquiring 20 machines. *The Star*.

Schoenbaum, M., & McCarty, D. (2019). Employment status and its impact on health outcomes: Evidence from chronic disease studies. *Journal of Health Economics, 68*, 102-114.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schroeder, E. B., Yang, X., Thorp, M. L., Arnold, B. M., Tabano, D. C., Petrik, A. F., Smith, D. H., Platt, R. W., & Johnson, E. S. (2017). Predicting 5-Year Risk of RRT in Stage 3 or 4 CKD: Development and External Validation. *Clinical journal of the American Society of Nephrology: CJASN*, *12*(1), 87–94.

Selby, N. M., & Taal, M. W. (2020). An updated overview of diabetic nephropathy: Diagnosis, prognosis, treatment goals and latest guidelines. *Diabetes, Obesity and Metabolism*, *22*, 3-15.

Selingerova, I., Katina, S., & Horova, I. (2021). Comparison of parametric and semiparametric survival regression models with kernel estimation. *Journal of Statistical Computation and Simulation*, *91*(13), 2717–2739.

Shavlik, J. W., & Dietterich, T. G. (1990). Readings in Machine Learning. In *Google Books*.

Shen, Y., Cai, R., Sun, J., Dong, X., Huang, R., Tian, S., & Wang, S. (2017). Diabetes mellitus as a risk factor for incident chronic kidney disease and end-stage renal disease in women compared with men: a systematic review and meta-analysis. *Endocrine*, *55*, 66-76.

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007, October 1). *A Support Vector Approach to Censored Targets*. IEEE Xplore.

Smith, H., Sweeting, M., Morris, T., & Crowther, M. J. (2022). A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagnostic and Prognostic Research*, *6*(1), 10.

Staub, L., & Gekenidis, A. (2011). "Kaplan–Meier Survival Curves and the Log-Rank Test" (PDF). *Survival Analysis (PDF). Handout and presentation. Seminar for Statistics (SfS).* Eidgenössische Technische Hochschule Zürich (ETH) [Swiss Federal Institute of Technology Zurich].

Stengel, B., & Pueyo, P. (2017). Tobacco smoking and kidney disease: A review. *Kidney International, 91*(2), 225-236.

Stenholm, S., & Head, J. (2018). Marital status and survival among diabetic patients: Evidence from the UK Biobank study. *Diabetes Care, 41*(7), 1451-1458.

Tapp, R. J., Shaw, J. E., Zimmet, P. Z., Balkau, B., Chadban, S. J., Tonkin, A. M., Welborn, T. A., & Atkins, R. C. (2004). Albuminuria is evident in the early stages of diabetes onset: results from the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). *American journal of kidney diseases: the official journal of the National Kidney Foundation*, *44*(5), 792–798.

Tekalign, T., Guta, M. T., Awoke, N., Chichiabellu, T. Y., Meskele, M., Anteneh, G., Tura, T. S., & Workie, S. B. (2023). Time to Diabetic Nephropathy and its Predictors Among Diabetic Patients Treated in Wolaita and Dawuro Zone Hospitals, Ethiopia: A Retrospective Cohort Study. *International journal of nephrology and renovascular disease*, *16*, 163–172.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Therneau, t. M., & li, h. (1999). Computing the cox model for case cohort designs. *Lifetime data analysis*, *5*, 99-112.

Thomas, M. C., Brownlee, M., Susztak, K., Sharma, K., Jandeleit-Dahm, K. A., Zoungas, S., & Cooper, M. E. (2015). Diabetic kidney disease. *Nature reviews Disease primers*, *1*(1), 1-20.

Thomas, N., Elliott, E. J., & Naughton, G. A. (2006). Exercise for type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews, 3*, CD002968.

Tong, R., Zhu, Z., & Ling, J. (2023). Comparison of linear and non-linear machine learning models for time-dependent readmission or mortality prediction among hospitalized heart failure patients. *Heliyon*, *9*(5), e16068.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.

Tuckman, B. W. (1972). Conducting Educational Research. In *Google Books*. Harcourt BraceJovanovich.

Tuttle, K. R., Bakris, G. L., Bilous, R. W., Chiang, J. L., de Boer, I. H., Goldstein-Fuchs, J., Hirsch, I. B., Kalantar-Zadeh, K., Narva, A. S., Navaneethan, S. D., Neumiller, J. J., Patel, U. D., Ratner, R. E., Whaley-Connell, A. T., & Molitch, M. E. (2014). Diabetic Kidney Disease: A Report from an ADA Consensus Conference. *Diabetes care*, *37*(10), 2864–2883.

Umanath, K., & Lewis, J. B. (2018). Update on Diabetic Nephropathy: Core Curriculum 2018. *American journal of kidney diseases: the official journal of the National Kidney Foundation*, *71*(6), 884–895.

Uusitupa, M. (2002). Lifestyles Matter in the Prevention of Type 2 Diabetes. *Diabetes Care*, *25*(9), 1650–1651.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., & Zhao, S. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nature reviews Drug discovery*, *18*(6), 463-477.

Van Belle, V., Pelckmans, K., Suykens, J., & Van Huffel. S., (2007). *Support vector machines for survival analysis.* In Proceedings of the Third International Conference on Computational Intelligence in Medicineand Healthcare (CIMED2007), pages 1–8, [p413, 414, 415]

Van Belle, V., Pelckmans, K., Suykens, J., & Van Huffel. S., (2007). *Support vector machines for survival analysis.* In Proceedings of the Third International Conference on Computational Intelligence in Medicineand Healthcare (CIMED2007), pages 1–8, [p413, 414, 415]

Van Belle, V., Pelckmans, K., Suykens, J., & Van Huffel.S., (2008) *Survival SVM : a practical scalable algorithm.* In European Symposium on Artificial Neural Networks (ESANN), pages 89–94, [p413, 414, 420]

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, *53*(2), 107–118.

Van Houwelingen, H. C., & Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Vapnik, V.N., *Statistical Learning Theory*. John Wiley & Sons, New York, (1998). ISBN 978-0-471-03003-4. [p413, 414]

Vapnik,V.J. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag,. ISBN 978-0-387-94559-0. [p412, 414]

Varghese, R.T., & Jialal, I. (2023, July 24). *Diabetic nephropathy - StatPearls - NCBI bookshelf*. National Center for Biotechnology Information.

Vart, P., & Li, M. (2018). The Interplay between Cardiovascular Diseases and Diabetic Kidney Disease: Mechanisms and Clinical Implications. *Diabetes Care, 41*(6), 1180-1190.

Verma, A., Vyas, S., Agarwal, A., Abbas, S., Agarwal, D. P., & Kumar, R. (2016). Diabetic Kidney Disease and Hypertension: A True Love Story. *Journal of clinical and diagnostic research: JCDR*, *10*(3), OC11–OC13.

Walker, R. J., Garacci, E., Campbell, J. A., Harris, M., Mosley-Johnson, E., & Egede, L. E. (2021). Relationship Between Multiple Measures of Financial Hardship and Glycemic Control in Older Adults With Diabetes. *Journal of applied gerontology : the official journal of the Southern Gerontological Society*, *40*(2), 162–169.

Walker, R. J., Garacci, E., Palatnik, A., Ozieh, M. N., & Egede, L. E. (2020). The longitudinal influence of social determinants of health on glycemic control in elderly adults with diabetes. *Diabetes Care*, *43*(4), 759-766.

Wang, J., Chen, Y., Xu, W., Lu, N., Cao, J., & Yu, S. (2019). Effects of intensive blood pressure lowering on mortality and cardiovascular and renal outcomes in type 2 diabetic patients: a meta-analysis. *PLoS One*, *14*(4), e0215362.

Weldegiorgis, M., Smith, M., Herrington, W. G., Bankhead, C., & Woodward, M. (2020). Socioeconomic disadvantage and the risk of advanced chronic kidney disease: results from a cohort study with 1.4 million participants. *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association*, *35*(9), 1562–1570.

Wens J., Vermeire E., Royen P. V., Sabbe B. & Denekens, 2005. Perspectives of type 2 diabetes patients‟ adherence to treatment: A qualitative analysis of barriers and solutions. *BMC Family Practice, 6: 20 -29.*

Willers, C., Iderberg, H., Axelsen, M., Dahlström, T., Julin, B., Leksell, J., Lindberg, A., Lindgren, P., Looström Muth, K., Svensson, A.-M., & Lilja, M. (2018). Sociodemographic determinants and health outcome variation in individuals with type 1 diabetes mellitus: A register-based study. *PLOS ONE*, *13*(6), e0199170.

Williams, A. J., & Hall, J. L. (2021). Enhancing Predictive Models for CKD by Including Family History: Implications for Clinical Practice. *Journal of Renal Care, 47*(3), 123-130.

Wolf, G., Busch, M., Müller, N., & Müller, U. A. (2011). Association between socioeconomic status and renal function in a population of German patients with diabetic nephropathy treated at a tertiary centre. *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association*, *26*(12), 4017–4023.

Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y., & Zheng, Y. (2022). The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study. *JMIR Medical Informatics*, *10*(2), e33440.

Xie, J., & Liu, C. (2005). Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in medicine*, *24*(20), 3089-3110.

Yirsaw, B. D. (2012). Chronic kidney disease in sub-Saharan Africa: Hypothesis for research demand. *Annals of African medicine*, *11*(2), 119-120.

Zhang, H., & Zhao, Y. (2017). Hybrid Models for Data Classification and Regression. *In Advances in Data Classification and Analysis*. Springer.

Zhang, H., Xia, W., Lu, X., Sun, R., Wang, L., Zheng, L., Ye, Y., Bao, Y., Xiang, Y., & Guo, X. (2013). A novel statistical prognostic score model that includes serum CXCL5 levels and clinical classification predicts risk of disease progression and survival of nasopharyngeal carcinoma patients. *PloS one*, *8*(2), e57830.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., & Gaciong, Z. (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *The lancet*, *387*(10027), 1513-1530.

APPENDICES

Appendix I: Questionnaire

(TO BE FILLED BY DIABETIC PATIENTS)

Date…….

Introduction: The study involves developing a model to estimate time to diabetic kidney disease in diabetic patients, and its predictors. Please take a few minutes to fill in this questionnaire

Please respond to all questions. Do not write your names.

Section 1: Demographic information (*Tick appropriately*)

1. Age

   (a) 18-25 [ ] (b) 26-35 [ ] (c) 36-45 [ ] (d) 46-55 [ ] (e) 56-65 [ ] (f) Above 65 [ ]

2. Gender

   Male [ ] Female[ ]

3. Marital status
   a. Never married [ ]
   b. Separated [ ]
   c. Divorced [ ]
   d. Widowed [ ]
   e. Married [ ]
4. What is the highest level of education you have completed?
   a. Primary school and below [ ]
   b. Secondary School [ ]
   c. Tertiary level [ ]
5. Current work situation
   a. Working full time [ ]
   b. Working part time [ ]
   c. Not working and not looking for work [ ]
   d. Unemployed and looking for work [ ]
   e. Disabled or retired and not looking for work [ ]
   f. Currently in School [ ]

Section 2: Lifestyle and Behaviour (*tick appropriately)*

6. Smoking Status:
   - [ ] Never smoked
   - [ ] Former smoker
   - [ ] Current smoker
7. Alcohol Consumption:
   - [ ] None
   - [ ] Occasional (less than once a week)
   - [ ] Moderate (1-3 times a week)
   - [ ] Frequent (4 or more times a week)
8. Current Weight _____ Kgs

Section 3: Financial Status (*tick appropriately)*

9. How do you pay for your healthcare and medical expenses?
   a. NHIF [ ]
   b. Private Insurance [ ]
   c. Self-Pay, out of pocket [ ]
10. Please read through the following statements and tick appropriately about how you have been able to cope in financial hardships KEY: 1. Strongly Agree 2. Agree 3. Undecided 4. Disagree 5. Strongly Disagree

Statements on Financial hardships	**1**	**2**	**3**	**4**	**5**
It has been difficult for me/my family to meet payments on my/my family bills					
I sometimes eat less than I should because I do not have enough money for food					
I take less medication than what is prescribed for me because of their cost					

Section 4: Medical History (*tick appropriately)*

11. How long have you had Diabetes? [ ] years and/or [ ]months

12. Other Medical Conditions (check all that apply):

- [ ] Hypertension
- [ ] Heart disease
- [ ] Kidney disease If YES, how long have you had kidney problems? [ ] years and/or [.] months
- [ ] Other (please specify): _____

13. Is there a past history of kidney disease in your household? Yes [ ]/ No [ ]

THANK YOU SO MUCH FOR YOUR PARTICIPATION.

Appendix II: Secondary Data Checklist

1. Socio- Demographic Background
    a) Patient's age [ ]
    b) Gender [ ]
    c) Weight [ ]
    d) BMI [ ]
    e) Marital Status [ ]
    f) Level of Education [ ]
    g) Current work situation [ ]
2. Medical History
    a) Date patient was diagnosed with Diabetes [ ]
    b) Blood sugar levels (Hemoglobin A1c) [ ]
    c) Blood pressure levels [ ]
    d) Cardio-vascular disease [ ]
    e) Patients' creatinine levels [ ]
    f) History of kidney disease in patients' household [ ]
3. Lifestyle and behavior
    a) Alcohol use [ ]
    b) Tobacco use [ ]
4. Financial Status
    a) How patient pay for healthcare [ ]
    b) If patient is able to acquire all medications prescribed [ ]

Appendix III: R CODES

```
#Cox Regression Model for Modelling DKD
# Loading required packages
library(survival) # for core survival analysis routines
library(psych) # for description of data
library(survminer) # drawing survival curves using 'ggplot2'
library(SurvMetrics) # predictive evaluation metrics in survival analysis
library(pec) # validation of risk predictions obtained from survival models
library(ggplot2) # create elegant data visualisations
library(ggpubr) # creating and customizing 'ggplot2'- based publication ready plots.
library(tidyverse) #
library(tibble) #
library(knitr) #
library(dplyr) #
library(rms) # regression modeling, testing, estimation, validation, graphics, prediction,
and typesetting

#Loading data into R
data1<-read.csv("C:/Users/Armstrong/Diabdata.csv",header = T)
#Exploratory of the dataset
dim(data1)
names(data1)     # View the variable/features/column names
str(data1) #Descriptive Statistics
summary(data1) #Descriptive Statistics
describe (data1) # Descriptive Statistics

#changing character variables to factors(categorical variables)
data1[sapply(data1,is.character)]<-lapply(data1[sapply(data1,is.character)],as.factor)
lapply(data1$Time,as.numeric)
lapply(data1$Diabeticage,as.numeric)
```

```
# Creating boxplots for visual representation of data
boxplot(Diabeticage~Event,data=data1,
        xlab = "Status",ylab = "Time in Years",
        main="Effect of Age",
        lwd=2,#thickness
        border="blue",#color of box borders,
        col="slategray2",#color inside the boxes
        frame=FALSE,
        boxwex=0.5)
boxplot(Weight~Event,data=data1,
        xlab = "Status",ylab = "Weight in Kilograms",
        main="Effect of Weight",
        lwd=2,#thickness
        border="blue",#color of box borders,
        col="slategray2",#color inside the boxes
        frame=FALSE,
        boxwex=0.5)
boxplot(Time~Event,data=data1,
        xlab = "Event",ylab = "Time in Years",
        main="Analysis on Time to DKD",
        lwd=2,#thickness
        border="blue",#color of box borders,
        col="slategray2",#color inside the boxes
        frame=FALSE,
        boxwex=0.5)

#creating a Surv object, life table and K-M curve
km<-survfit(Surv(Time,Event)~-1,data = data1)
km
summary(km,times = seq(0,30)*3)
plot(km,xlab = "Time in Years",ylab = "Survival Function",
```

```
    main="K_M Survival Curve", mark.time = F, conf.int = T, col="blue")
ggsurvplot(km, palette = "blue",risk.table = TRUE,surv.median.line = c("hv"))
#mean survival time
print(km,print.rmean = TRUE)

#Performing log-rank tests
#Procedure for each involves
#The following hypotheses are used in this test:
#H0: There is no difference in survival between the groups.
#HA: There is a difference in survival between the groups.
lrtGender<-survdiff(Surv(Time,Event)~Gender,data = data1)
lrtGender
kmG<-survfit(Surv(Time,Event)~Gender,data=data1)
ggsurvplot(kmG,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmG, rmean = 30)

lrtHTN<-survdiff(Surv(Time,Event)~HTN,data = data1)
lrtHTN
kmH<-survfit(Surv(Time,Event)~HTN,data=data1)
ggsurvplot(kmH,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmH, rmean = 31)

lrtCVD<-survdiff(Surv(Time,Event)~CVD,data = data1)
lrtCVD
kmC<-survfit(Surv(Time,Event)~CVD,data=data1)
ggsurvplot(kmC,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmC, rmean = 30)

# Age at diabetes
lrtDiabeticage<-survdiff(Surv(Time,Event)~Diabeticage,data=data1)
lrtDiabeticage# we need to check for linearity
```

```
coxd<-coxph(Surv(Time,Event)~Diabeticage,data=data1)
summary(coxd)
Resid<-residuals(coxd,type = "martingale")
plot(data1$Diabeticage,Resid,xlab ="Age at Diabetes Diagnosis",
    ylab = "Martingale Residuals")
testph1=cox.zph(coxd)
testph1
ggcoxzph(testph1)
ggcoxfunctional(Surv(Time,Event)~Diabeticage+log(Diabeticage)+sqrt(Diabeticage),dat
a=data1)

lrtWeight<-survdiff(Surv(Time,Event)~Weight,data = data1)
lrtWeight
coxw<-coxph(Surv(Time,Event)~Weight,data=data1)
summary(coxw)
Resid1<-residuals(coxw,type = "martingale")
plot(data1$Diabeticage,Resid1,xlab ="Weight at Diabetes Diagnosis",
    ylab = "Martingale Residuals")
testph=cox.zph(coxw)
plot(testph)
testph
coxw
ggcoxzph(testph)

lrtEducation<-survdiff(Surv(Time,Event)~Education,data = data1)
lrtEducation
kmE<-survfit(Surv(Time,Event)~Education,data=data1)
ggsurvplot(kmE,data = data1,risk.table = F,pval = T,conf.int = F)
survival:::survmean(kmE, rmean = 30)

lrtMaritalstatus<-survdiff(Surv(Time,Event)~Spouse,data = data1)
```

```
lrtMaritalstatus
kmM<-survfit(Surv(Time,Event)~Spouse,data=data1)
ggsurvplot(kmM,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmM, rmean = 30)

lrtTobacco<-survdiff(Surv(Time,Event)~Tobacco,data = data1)
lrtTobacco
kmT<-survfit(Surv(Time,Event)~Tobacco,data=data1)
ggsurvplot(kmT,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmT, rmean = 30)

lrtAlcohol<-survdiff(Surv(Time,Event)~Alcohol,data = data1)
lrtAlcohol
kmA<-survfit(Surv(Time,Event)~Alcohol,data=data1)
ggsurvplot(kmA,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmA, rmean = 30)

lrtHistory<-survdiff(Surv(Time,Event)~History,data = data1)
lrtHistory
kmHi<-survfit(Surv(Time,Event)~History,data=data1)
ggsurvplot(kmHi,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmHi, rmean = 30)

lrtExercise<-survdiff(Surv(Time,Event)~Exercise,data = data1)
lrtExercise
kmEx<-survfit(Surv(Time,Event)~Exercise,data=data1)
ggsurvplot(kmEx,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmEx, rmean = 30)

lrtFinancialhardship<-survdiff(Surv(Time,Event)~Financialhardship,data = data1)
lrtFinancialhardship
```

```
kmF<-survfit(Surv(Time,Event)~Financialhardship,data=data1)
ggsurvplot(kmF,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmF, rmean = 30)

lrtEmployment<-survdiff(Surv(Time,Event)~Employment,data = data1)
lrtEmployment
kmEmp<-survfit(Surv(Time,Event)~Employment,data=data1)
ggsurvplot(kmEmp,data = data1,risk.table = F,pval = TRUE,conf.int = F)
survival:::survmean(kmEmp, rmean = 30)

# cox model
library(StepReg)
x<-stepwise(formula=Surv(Time,Event)~.,data = data1,type = "cox", include =
c("Financialhardship"),metric = "AIC")
x

fit_x<-coxph(Surv(Time,Event)~Diabeticage+Gender+Employment+History+
        Financialhardship+Education+Alcohol,data=data1)
fit_x
summary(fit_x)
anova(fit_x)
cox.zph(fit_x)
concordance(fit_x)
extractAIC(fit_x)

#Adjusted data
data2<-read.csv("C:/Users/Armstrong/Desktop/Diabdat.csv",header = T)
#splitting data
set.seed(123)
index = sample(2, nrow(data2),replace =T, prob=c(0.70,0.30))
traindata2 = data2[index ==1,]
```

```
testdata2 = data2[index ==2,]

fit_x_adj<-coxph(Surv(Time,Event)~Diabeticage+Gender+AdjEmp+History+
          Financialhardship+AdjEdu+Alcohol,data=traindata2)
fit_x_adj
summary(fit_x_adj)
x<-cox.zph(fit_x_adj)
x
p<-survfit(fit_x_adj)
plot(p)
ggcoxzph(x)
ggforest(fit_x_adj)
anova(fit_x_adj)
concordance(fit_x_adj)
extractAIC(fit_x_adj)

y<-survival::Surv(testdata$Time,testdata$Event)
x<-predict(fit_x,newdata = testdata2,type = "survival")
concordance(x)
summary(x)
testdata
survfit(x)

# Survival SVM
## This section will include:
## 1) libraries to be used are loaded,
## 2) required learners are wrapped in the R package mlr,
## 3) measures are defined,
# Load required libraries
library(survivalsvm)
library(mlr)
```

```
library(caret)
library(Hmisc)
library(Rmisc)
library(BBmisc)
library(checkmate)
library(survival)
library(ggplot2)
library(mlr)
library(plyr)
# Wrap learners into the mlr package

#--- creation of an mlr leaner for survivalsvm
makeRLearner.surv.survivalsvm = function() {
  makeRLearnerSurv(
    cl = "surv.survivalsvm",
    package = "survivalsvm",
    par.set = makeParamSet(
      makeDiscreteLearnerParam(id = "type",  default = "regression",
                               values = c("regression", "vanbelle1", "vanbelle2",
                                          "hybrid")),
      makeDiscreteLearnerParam(id = "diff.meth",  default = "makediff3",
                               values = c("makediff1", "makediff2",
                                          "makediff3")),
      makeNumericVectorLearnerParam(id = "gamma.mu",
                                    tunable = TRUE, lower = 2^-2, upper = 2^2),
      makeDiscreteLearnerParam(id = "opt.meth", default = "quadprog",
                               values = c("quadprog", "ipop")),
      makeDiscreteLearnerParam(id = "kernel", default = "lin_kernel",
                               values = c("lin_kernel", "add_kernel",
                                          "rbf_kernel",
                                          "rbf4_kernel", "poly_kernel")),
```

```
    makeNumericLearnerParam(id = "kernel.pars", tunable = TRUE),
    makeNumericLearnerParam(id = "sgf.sv", default = 5, tunable = FALSE),
    makeNumericLearnerParam(id = "sigf", default = 7, tunable = FALSE),
    makeNumericLearnerParam(id = "maxiter", default = 20, tunable = FALSE),
    makeNumericLearnerParam(id = "margin", default = 0.05, tunable = FALSE),
    makeNumericLearnerParam(id = "bound", default = 10, tunable = FALSE),
    makeNumericLearnerParam(id = "eig.tol", default = 1e-06,
                    tunable = FALSE),
    makeNumericLearnerParam(id = "conv.tol", default = 1e-07,
                    tunable = FALSE),
    makeNumericLearnerParam(id = "posd.tol", default = 1e-08,
                    tunable = FALSE) ),
  properties = c("missings", "numerics", "factors", "weights", "prob",
            "rcens"),
  name = "survival support vector machines",
  short.name = "survivalsvm",
  note = "survivalsvm in mlr")}

#--- creation of trainer for survivalsvm
trainLearner.surv.survivalsvm = function(.learner, .task, .subset, ...) {
 f <-  getTaskFormula(.task)
 data  <-  getTaskData(.task, subset = .subset)
 mod <-  survivalsvm::survivalsvm(formula = f, data = data, ...)
 return(mod)}

#--- creation of predictor for survivalsvm
predictLearner.surv.survivalsvm = function(.learner, .model, .newdata, ...) {
 if (.learner$predict.type == "response") {
  predict(object = .model$learner.model,
       newdata = .newdata, ...)$predicted[1,]}}
```

```
#--- Wrapper for scale data when required
makePreprocWrapperScale = function(learner, center = TRUE, scale = TRUE) {
 trainfun = function(data, target, args = list(center, scale)) {
  cns = colnames(data)
  nums = setdiff(cns[sapply(data, is.numeric)], target)
  x = as.matrix(data[, nums, drop = FALSE])
  x = scale(x, center = args$center, scale = args$scale)
  control = args
  if (is.logical(control$center) && control$center)
   control$center = attr(x, "scaled:center")
  if (is.logical(control$scale) && control$scale)
   control$scale = attr(x, "scaled:scale")
  data = data[, setdiff(cns, nums), drop = FALSE]
  data = cbind(data, as.data.frame(x))
  return(list(data = data, control = control))}
 predictfun = function(data, target, args, control) {
  cns = colnames(data)
  nums = cns[sapply(data, is.numeric)]
  x = as.matrix(data[, nums, drop = FALSE])
  x = scale(x, center = control$center, scale = control$scale)
  data = data[, setdiff(cns, nums), drop = FALSE]
  data = cbind(data, as.data.frame(x))
  return(data)}
 if(!("surv.glmboost" %in% class(learner)))
  makePreprocWrapper(
   learner,
   train = trainfun,
   predict = predictfun,
   par.set = makeParamSet(
    makeLogicalLearnerParam("center"),
    makeLogicalLearnerParam("scale")),
```

```
    par.vals = list(center = center, scale = scale))
  else
    makePreprocWrapper(
      learner,
      train = trainfun,
      predict = predictfun,
      par.set = makeParamSet(),
      par.vals = list(center = center, scale = scale))}
# --- TuneWrapper for survivalsvm
tuneWrapperGammaMu <- function(type, kernel, ...,
                               preproc = TRUE, tune.scale = TRUE, center = TRUE,
                               scale = TRUE, resolution = 5L, method = "CV",
                               lower = -5L, upper = 5L, iters.rep = 4L) {
  lrn <- makePreprocWrapperScale(makeLearner(cl = "surv.survivalsvm",
                                             type = type, kernel = kernel,
                                             ...),
                                 center = center, scale = scale)
  configureMlr(on.learner.error = "warn")
  if(kernel %in% c("lin_kernel", "add_kernel")) {
    # Parameters set for linear and additive kernel
    if (type != "hybrid") {
      discrete_ps <- if(tune.scale) {
        makeParamSet(
          makeDiscreteParam("gamma.mu", values = 2^(lower:upper)),
          makeLogicalLearnerParam("center"),
          makeLogicalLearnerParam("scale"))} else {
        makeParamSet(
          makeDiscreteParam("gamma.mu",
                            values = 2^(lower:upper)) )}} else {
      discrete_ps <- if(tune.scale) {
        makeParamSet(
```

```
    makeNumericVectorParam("gamma.mu", len = 2L, lower = lower,
                           upper = upper,
                           trafo = function(x) 2^x),
    makeLogicalLearnerParam("center"),
    makeLogicalLearnerParam("scale"))} else {
   makeParamSet(
    makeNumericVectorParam("gamma.mu", len = 2L, lower = lower,
                           upper = upper,
                           trafo = function(x) 2^x))}}} else {
# Parameters set for RBF kernel
if (type != "hybrid") {
 discrete_ps <- if(tune.scale) {
   makeParamSet(
    makeDiscreteParam("gamma.mu", values = 2^(lower:upper)),
    makeDiscreteParam("kernel.pars", values = 2^(-5:5)),
    makeLogicalLearnerParam("center"),
    makeLogicalLearnerParam("scale"))} else {
   makeParamSet(
    makeDiscreteParam("gamma.mu",
                      values = 2^(lower:upper)),
    makeDiscreteParam("kernel.pars", values = 2^(-5:5)))}} else {
 discrete_ps <- if(tune.scale) {
   makeParamSet(
    makeNumericVectorParam("gamma.mu", len = 2L, lower = lower,
                           upper = upper,
                           trafo = function(x) 2^x),
    makeLogicalLearnerParam("center"),
    makeLogicalLearnerParam("scale"),
    makeDiscreteParam("kernel.pars", values = 2^(-5:5)))} else {
   makeParamSet(
    makeNumericVectorParam("gamma.mu", len = 2L, lower = lower,
```

```
                        upper = upper,
                        trafo = function(x) 2^x),
        makeDiscreteParam("kernel.pars", values = 2^(-5:5)))}}}
  ctrl  <-  makeTuneControlGrid(resolution = resolution)
  inner  <-  makeResampleDesc(method = method, iters = iters.rep)
  survivalsvm.tuned  <-  makeTuneWrapper(lrn,  resampling = inner,
                                 par.set = discrete_ps, control = ctrl,
                                 measures = c.i)
  return(survivalsvm.tuned)}
# ------------------------------------------------------------------------------
# Wrap measures into the mlr package
# ------------------------------------------------------------------------------
my.ci.fun <- function(task, model, pred, feats, extra.args) {
  myci = rcorr.cens(x = getPredictionResponse(pred),
                S = getPredictionTruth(pred))
  return(myci["C Index"])}
# constructs the C-index measure for survivalsvmprediction objects.
c.i <- makeMeasure(
  id = "ci", name = "C-Index",
  properties = c("surv"),
  minimize = FALSE, best = 1, worst = 0,
  fun = my.ci.fun)
# ------------------------------------------------------------------------------
# Benchmarking for the diabdata dataset
# ------------------------------------------------------------------------------
# #Some pre-processing procedures
set.seed(152)
data3<-read.csv("C:/Users/Armstrong/Desktop/Diabdata.csv",header = T)
data3[sapply(data3,is.character)]<-lapply(data3[sapply(data3,is.character)],as.factor)
data.adj <- data3
data.task <- makeSurvTask(data = data.adj, target = c("Time", "Event"))
```

```
outer <- makeResampleDesc("CV", iters = 2L)

# Linear Kernel #
# --- survivalsvm regression.tuned
set.seed(123)
tunwrp.gm.reg <- tuneWrapperGammaMu(type = "regression",
                                    kernel = "lin_kernel",
                                    opt.meth = "quadprog", method = "CV",
                                    iters.rep = 3)
bench.dat.ssvm.reg <- benchmark(learners = tunwrp.gm.reg, tasks = data.task,
                                resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.reg$results$data.adj$surv.survivalsvm.preproc.tuned$
       measures.test, function(i){
         a <- CI(i)
         err <- a[1] - a[2]
         mittelwert <- a[2]
         r <- round(c(mittelwert, err), 2)
         names(r) <- c("mean", "error")
         return(r)})
# --- survivalsvm hybrid.tuned
set.seed(123)
tunwrp.gm.hyb <- tuneWrapperGammaMu(type = "hybrid",
                                    kernel = "lin_kernel",
                                    opt.meth = "quadprog",
                                    diff.meth = "makediff3")
bench.dat.ssvm.hyb <- benchmark(learners = tunwrp.gm.hyb, tasks = data.task,
                                resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.hyb$results$data.adj$surv.survivalsvm.preproc.tuned$
       measures.test, function(i){
         a <- CI(i)
         err <- a[1] - a[2]
```

```
        mittelwert <- a[2]
        r <- round(c(mittelwert, err), 2)
        names(r) <- c("mean", "error")
        return(r)})

# additive Kernel #
# --- survivalsvm regression.tuned
set.seed(123)
tunwrp.gm.ak.reg <- tuneWrapperGammaMu(type = "regression",
                              kernel = "add_kernel",
                              opt.meth = "quadprog", method = "CV",
                              iters.rep = 5)
bench.dat.ssvm.ak.reg <- benchmark(learners = tunwrp.gm.ak.reg,
                           tasks = data.task,
                           resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.ak.reg$results$data.adj$surv.survivalsvm.preproc.tuned$
       measures.test, function(i){
        a <- CI(i)
        err <- a[1] - a[2]
        mittelwert <- a[2]
        r <- round(c(mittelwert, err), 2)
        names(r) <- c("mean", "error")
        return(r)})
# --- survivalsvm hybrid.tuned
set.seed(123)
tunwrp.gm.Ak.hyb <- tuneWrapperGammaMu(type = "hybrid",
                              kernel = "add_kernel",
                              opt.meth = "quadprog",
                              diff.meth = "makediff3")
bench.dat.ssvm.ak.hyb <- benchmark(learners = tunwrp.gm.Ak.hyb,
                           tasks = data.task,
```

```
                              resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.ak.hyb$results$data.adj$surv.survivalsvm.preproc.tuned$
       measures.test, function(i){
         a <- CI(i)
         err <- a[1] - a[2]
         mittelwert <- a[2]
         r <- round(c(mittelwert, err), 2)
         names(r) <- c("mean", "error")
         return(r)})

# RBF/Gaussian kernel #
# --- survivalsvm regression.tuned
set.seed(123)
tunwrp.gm.rbf.reg <- tuneWrapperGammaMu(type = "regression",
                                kernel = "rbf_kernel",
                                opt.meth = "quadprog", method = "CV",
                                iters.rep = 5)
bench.dat.ssvm.rbf.reg <- benchmark(learners = tunwrp.gm.rbf.reg,
                            tasks = data.task,
                            resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.rbf.reg$results$data.adj$surv.survivalsvm.preproc.tuned$
       measures.test, function(i){
         a <- CI(i)
         err <- a[1] - a[2]
         mittelwert <- a[2]
         r <- round(c(mittelwert, err), 2)
         names(r) <- c("mean", "error")
         return(r)})
# --- survivalsvm hybrid.tuned
set.seed(123)
tunwrp.gm.rbf.hyb <- tuneWrapperGammaMu(type = "hybrid",
```

```
                                kernel = "rbf_kernel",
                                opt.meth = "quadprog",
                                diff.meth = "makediff3")
bench.dat.ssvm.rbf.hyb <- benchmark(learners = tunwrp.gm.rbf.hyb,
                                tasks = data.task,
                                resamplings = outer, measures = list(c.i))
lapply(bench.dat.ssvm.rbf.hyb$results$data.adj$surv.survivalsvm.preproc.tuned$
      measures.test, function(i){
        a <- CI(i)
        err <- a[1] - a[2]
        mittelwert <- a[2]
        r <- round(c(mittelwert, err), 2)
        names(r) <- c("mean", "error")
        return(r)})
```

Appendix IV: Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 758013

Date of Issue: 04/July/2024

RESEARCH LICENSE

This is to Certify that Miss.. GRACE MAKENA NJOKA of Chuka University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Embu, Kirinyaga, Meru on the topic: MODELLING OF PREDICTORS OF DIABETIC KIDNEY DISEASE AMONG DIABETIC PATIENTS for the period ending : 04/July/2025.

License No: **NACOSTI/P/24/37257**

758013

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions