# Methodology

## 1.1. Data Source, Limitations and Improvement Opportunities

The data used for this analysis were obtained from a local healthcare provider and consisted of a single, randomly selected primary care visit for each patient over ten years. However, it is important to note that this dataset did not capture variations in certain time-varying factors such as BMI, which can change as a person ages. To account for this variation, having access to an individual's entire primary visit hospital data records over the ten years would have helped understand how these factors affected the outcome of cardiovascular disease in five years. With access to this data, we could have used a range of variables that have been reported to affect the outcome of cardiovascular diseases, such as medication history(Okshina et al., 2019). This would have allowed us to use multi-level modelling to better comprehend the random effects of these factors on the outcome.

## 1.2. Outcome

The outcome of interest in this study was the occurrence of cardiovascular disease (CVD), denoted as status, a binary indicator representing whether a patient received a diagnosis of CVD or not. The analysis utilized the recorded year (months were estimated to years) of either the diagnosis of cardiovascular disease or the point of censoring, marking the conclusion of the observation period, to determine the outcome year.

## 1.3. Predictors

The predictor variables in the dataset included relevant information on patients, such as age, sex, BMI, smoking status, diabetes history, chronic kidney disease diagnosis, and atrial fibrillation diagnosis. These are described in the table below.

Table 1: Description of predictor variables

|  |  |  |
| --- | --- | --- |
| Variable name | Variable description | Variable data type and coding |
| Age | Age of the patient in years | Numeric: mean-centered |
| Sex | Sex of patient | Binary: (0 = female, 1 = male) |
| BMI | Body Mass Index of the patient at time 0 (weight in kg divided by height in meters squared) | Numeric: mean-centered |
| Smoking Status | Smoking status at time 0 | Binary: (0 = no, 1 = yes) |
| Diabetes | Diabetes diagnosis at time 0 | Binary: (0 = no, 1 = yes) |
| CKD | diagnosis of chronic kidney disease at time 0 | Binary: (0 = no, 1 = yes) |
| AF | Diagnosis of atrial fibrillation at time 0 | Binary: (0 = no, 1 = yes) |

## 1.4. Statistical Analysis Methods

### 1.4.1 Exploratory Data Analysis

Descriptive statistics were generated to provide a comprehensive overview of the dataset. The distribution of categorical variables, such as sex, smoking status, diabetes, chronic kidney disease, and atrial fibrillation, was summarized based on the occurrence of cardiovascular events. Additionally, histograms were constructed to visualize the distribution of continuous variables, including age and BMI.

### 1.4.2 Derivation and validation of the models

To investigate the relationship between the predictor variables and the time to cardiovascular events, the Cox Proportional Hazards model was created, validated and calibrated using established methodologies used to develop a similar clinical prediction model(Hippisley-Cox et al., 2017). This model assumes that the time-estimated hazard ratio is proportional to the length of time t and that the impact of each predictor in the model is constant in the relationship to the outcome for time t.(Hashim & Weiderpass, 2019) Initially, a simple model comprising basic demographic and clinical factors was considered however it violated the proportion hazards assumption. Univariate models for age and BMI and their restricted cubic spline transformation were created to test if non-linear associations with the time-to-event object were more significant than the normal values. We then fit a complex model that incorporated non-linear effects using restricted cubic splines for age and was stratified according to sex. Forward stepwise model selection based on the Akaike Information Criterion (AIC) was employed to assess the performance of complex models and to identify the most informative model with the least required predictors which was considered as the final model.

## 1.5. Validation of the model

Internal validation of the final model was performed using bootstrap resampling. The bootstrap method was employed for both model validation and calibration, with 200 bootstrap samples generated to assess the stability and performance of the final model. For validation, Harrell’s C statistics and R2 values were calculated. R2 values were calculated to determine the proportion of the variation in time to cardiovascular disease diagnosis we explained by the model, Harrell’s C statistic (equivalent to the area under the curve of a receiver operating curve) at 5 years was calculated to evaluate the discriminatory ability of the model to distinguish between patients with different event times. Calibration (comparing the mean predicted risks at five years with the observed risk corresponding to the 10th percentile of the predicted risk distribution). The observed risks were obtained using the Kaplan-Meier estimates evaluated at five years.

All analysis was implemented using R version 4.1. using the ‘rms’ package.(F. E. Harrell, 2015)

# 2.0. RESULTS

## 2.1. Exploratory Data Analysis for Predictor Variables

Table 2 shows the results of an in-depth analysis of predictor variables about the event of interest occurrence, with 4,415 individuals experiencing the event of interest and 10,947 classified as no event or censored. The distribution of sex revealed that 45% of the overall population was female, with a notable difference between those with an event (34% female) and those without (50% female). 3988 (26%) were either smokers or had they had smoked before with 11374(74%) being those who never smoked before displaying consistent proportions among both event and non-event groups. Regarding diabetes, 77% of the overall population was non-diabetic, while 23% were diabetic, indicating a shift towards higher diabetic proportions in the event subgroup. Chronic kidney disease exhibited a similar trend, with 68% of the total population without the condition and 32% with it, revealing a slightly elevated prevalence among those experiencing the event. Atrial fibrillation was present in 15% of the entire cohort, with a slightly higher prevalence among those with an event (17%).

Table 2: Summary Statistics for Predictor Variables

| **Predictor** | **Overall**, N = 15,362 | **event**, N = 4,415 | **no event/ censored**, N = 10,947 |
| --- | --- | --- | --- |
| sex |  |  |  |
| female | 6,971 (45) % | 1,493 (34) % | 5,478 (50) % |
| male | 8,391 (55) % | 2,922 (66) % | 5,469 (50) % |
| smoking status |  |  |  |
| non-smoker | 11,374 (74) % | 3,259 (74) % | 8,115 (74) % |
| smoker/  ever smoked | 3,988 (26) % | 1,156 (26) % | 2,832 (26) % |
| diabetic |  |  |  |
| no | 11,855 (77) % | 3,273 (74) % | 8,582 (78) % |
| yes | 3,507 (23) % | 1,142 (26) % | 2,365 (22) % |
| chronic\_kidney\_disease |  |  |  |
| no | 10,444 (68) % | 2,884 (65) % | 7,560 (69) % |
| yes | 4,918 (32) % | 1,531 (35) % | 3,387 (31) % |
| atrial\_fibrillation |  |  |  |
| no | 12,990 (85) % | 3,669 (83) % | 9,321 (85) % |
| yes | 2,372 (15) % | 746 (17) % | 1,626 (15) % |
|  |  |  |  |

The variables age and BMI in the dataset underwent mean centering. The distribution of BMI closely approximated a normal distribution, while the age variable exhibited a slight leftward skew, as illustrated in Figure 1 below.

A comparison of bmi distribution and bmi distribution

Description automatically generated

Figure : Distribution of Age and BMI

## 2.2. Model development: Simple Model vs Complex model

All the predictor variables were used to fit the simple Cox PH model. The proportional hazard assumption was tested on this simple model. The Schoenfeld residuals test (see table 3) on the simple model revealed a global X2 statistic of 22.8 with a p-value of 0.0018, indicating that the effect of the covariate was changing over time. To understand how the predictor's age and sex (p-values 0.3628, <0.00000 respectively) were affecting the outcome over time, their effect was individually investigated.

The graph shown in Figure 2 supports the finding that hazard proportions associated with sex changed over the ten years (X2 =4.52897, p-value =0.00002), which led to the decision to use a model stratified by sex(citation). The test for linearity of the association between age and BMI with the time-to-event object indicated that the model with a non-linear association of age and the time to event object was not significantly worse compared to model assuming age was linear (X2 = 96.26, p-value = 0.00000). However, the test for BMI (X2 = 0.0035, p-value = 0.95) indicated a slight difference in fit between the two models. Therefore, in the complex model, only a restricted cubic spline of age was included.

The Schoenfeld residuals test was conducted on the complex model (see table 3), revealed a global X2 statistic of 4.53 with a p-value of 0.717, indicating that the covariate effect was constant over 10 years. Therefore, the complex model was selected for further analysis.

Table 3: Proportional Hazard Tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Simple Model** | | **Complex model** | |
| **Variables** | **chq2** | **p-value** | **chq2** | **p-value** |
| age | 0.828 | 0.3628 | 0.998124 | 0.607 |
| sex | 18.3 | 2.00E-05 | strata |  |
| smoking status | 3.19 | 0.0741 | 3.112389 | 0.078 |
| bmi | 0.000101 | 0.992 | 0.000446 | 0.983 |
| diabetic | 0.319 | 0.5724 | 0.274313 | 0.6 |
| chronic\_kidney\_disease | 0.00175 | 0.9667 | 0.003412 | 0.953 |
| atrial\_fibrillation | 0.0112 | 0.9156 | 0.002507 | 0.96 |
| **Global** | **22.8** | **0.0018** | **4.52897** | **0.717** |
| *note: age in the complex model uses three restricted cubic splines* | | | | |

A graph with numbers and lines

Description automatically generated

Figure 2: Survival curve of sex over 10 years

## 2.3. Model specification

The forward stepwise Akaike Information Criterion (AIC) base model selection revealed that the reduced model with age (with a restricted cubic spline), BMI, diabetes and chronic kidney disease had the least AIC = 77431.94 (compared to AIC = 77432.88 in the initial complex model). The effect of the predictor variables on the outcome variable is shown in table 4 below:

Table 4: Adjusted Hazard ratios and their confidence intervals.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Adjusted HR** | **lower 95% C. I** | **upper 95% C. I** | **p-value** |
| age | 0.9601562 | 0.9483486 | 0.9704278 | <0.0001 |
| age' | 1.0665132 | 1.0548770 | 1.0813331 | <0.0001 |
| BMI | 1.0434512 | 1.0365650 | 1.0521276 | <0.0001 |
| Diabetic\_1 | 1.1866122 | 1.1141391 | 1.2750631 | <0.0001 |
| chronic\_kidney\_disease\_1 | 1.1119338 | 1.0474883 | 1.1864906 | <0.0001 |

For age, each one-unit increase corresponds to a 3.98% decrease in the hazard of experiencing the event, with an adjusted HR of 0.960 (95% C.I: 0.948, 0.970). The non-linear effect of age, represented by age', is associated with a 6.65% increase in the hazard for each unit change (adjusted HR: 1.067, 95% C.I: 1.055, 1.081). A higher BMI is linked to a 4.35% increase in the hazard for each one-unit increase (adjusted HR: 1.043, 95% C.I: 1.037, 1.052). Diabetic individuals exhibit an 18.66% higher hazard of cardiovascular events compared to non-diabetic individuals (adjusted HR: 1.187, 95% C.I: 1.114, 1.275). Moreover, individuals with chronic kidney disease experience an 11.19% higher hazard (adjusted HR: 1.112, 95% C.I: 1.047, 1.186).

## 2.4. Model Performance

### 2.4.1. Discrimination

The Discrimination Index (Dxy), a measure of the model's ability to distinguish between classes, yielded a value of 0.2036, indicating a low level of discriminatory power. The R-squared (R²) value of 0.0186 suggests that only 1.86% of the variance of the outcome variable was explained by the predictor variables in this model. This means most of the predictions from this model are most likely to be biased. Additionally, Harrell’s C statistics, which measures concordance, exhibited a value of 0.39, indicative of relatively low predictive accuracy.

### 2.4.2. Calibration

Based on the calibration plot (refer to Figure 3) of the Cox model, it appears that the model's predicted probabilities are well-calibrated, both before and after correcting for potential optimism. The observed and optimism-corrected lines are very close to the ideal 45-degree line (grey line), indicating good calibration. The mean absolute calibration error and quantile error are both very low (0.002), suggesting that the model's predictions are only off by 0.2% from the observed probabilities on average. These findings support the visual indication of good calibration, thus indicating that the predictive model is performing well in predicting the 5-year hazard of cardiovascular disease.

A graph with a blue line

Description automatically generated

Figure 3: Calibration Plot for Optimism reduction

## 2.4.5. Model Presentation

The baseline hazard for the final was calculated separately for males and females (because the model was stratified). For females, the baseline hazard at 5 years was found to be 0.1199 and for males, it was 0.1758. This information is shown in Table 5. Based on these values, the corresponding baseline survival probabilities were calculated to be 0.8914 for females and 0.8387 for males. The Cox PH model can be simplified by incorporating the baseline survival probabilities as a factor.

2

Table 5: Baseline hazard and survivor rates

|  |  |  |
| --- | --- | --- |
| **Strata** | **Baseline Hazard** | **Baseline Survival** |
| Sex = 0 | 0.1199 | 0.891376395 |
| Sex = 1 | 0.1758 | 0.838749992 |

## 2.4.6 Model Usage - Nomogram

The nomogram (see Figure 4 below) has been provided as a visual tool designed to estimate 5-year survival probabilities for cardiovascular disease through a Cox proportional hazards model(Iasonos et al., 2016). The key predictors in this model are age, BMI, diabetes status, and CKD. Each predictor is assigned points inversely based on age and BMI, indicating that the risk increases as age and BMI deviations increase. The points from all predictors are added to determine a total, which corresponds to a linear predictor value. This value is then converted to a 5-year survival probability, with the scale indicating that a higher total points value is associated with lower chances of survival.

A screenshot of a graph

Description automatically generated

Figure 4: Nomogram from model usage

Harrell, F. E. (2015). Introduction to Survival Analysis. In Jr. Harrell Frank E. (Ed.), *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (pp. 399–422). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7\_17

Hashim, D., & Weiderpass, E. (2019). Cancer Survival and Survivorship. In P. Boffetta & P. Hainaut (Eds.), *Encyclopedia of Cancer (Third Edition)* (pp. 250–259). Academic Press. https://doi.org/10.1016/B978-0-12-801238-3.65102-4

Hippisley-Cox, J., Coupland, C., & Brindle, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ*, *357*. https://doi.org/10.1136/bmj.j2099

Iasonos, A., Schrag, D., Raj, G. V., & Panageas, K. S. (2016). How To Build and Interpret a Nomogram for Cancer Prognosis. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2007.12.9791

Okshina, E. Y., Loukianov, M. M., Martsevich, S. Y., Yakushin, S. S., Kutishenko, N. P., Vorobyev, A. N., Pereverzeva, K. G., Zagrebelnyy, A. V., Voronina, V. P., Dmitrieva, N. A., Lerman, O. V., Kudryashov, E. V., Boytsov, S. A., & Drapkina, O. M. (2019). Patients with History of Myocardial Infarction and Acute Cerebrovascular Accidentin Clinical Practice: Demographic, Clinical Characteristics, Drug Treatment and Outcomes (Data of Outpatient and Hospital Registry REGION). *Rational Pharmacotherapy in Cardiology*, *15*(5), Article 5. https://doi.org/10.20996/1819-6446-2019-15-5-656-662

## Appendix- R Code used for the analysis.

# Setting Working Directory -----------------------------------------------

setwd("/Users/\*\*\*\*\*\*/Documents/MSc\_HDS/Stats\_Modelling/Assessment")

# Loading Libraries -------------------------------------------------------

library(tidyverse)

library(janitor)

library(survival)

library(gtsummary)

library(grid)

library(gridExtra)

library(rms)

library(stats)

library(survival)

library(conflicted)

library(DescTools)

library(survminer)

# Data Loading, Cleaning and Preparation ----------------------------------

cvd.df <- read\_csv("cvddata.csv") |>

mutate(

ID = as.character(ID),

Sex = as.factor(Sex),

Smoking\_Status = as.factor(Smoking\_Status),

Diabetic = as.factor(Diabetes),

CKD = as.factor(CKD),

atrial\_fibrillation = as.factor(AF),

TEVENT = TEVENT

)|>

select(-c(AF,ID,Diabetes)) |>

clean\_names() #use globally accepted valiable naming standards

# Summary characteristics --------------------------------------------------------

cvd.df |> select(

c(sex,smoking\_status,diabetic,status,ckd,atrial\_fibrillation)

) |> mutate(

sex = ifelse(sex == 0,"female","male"),

smoking\_status = ifelse(smoking\_status == 0,"no-smoker","smoker/ever\_smoked"),

diabetic = ifelse(diabetic == 0,"no","yes"),

chronic\_kidney\_disease = ifelse( ckd == 0,"no","yes"),

atrial\_fibrillation = ifelse(atrial\_fibrillation == 0,"no","yes"),

status = ifelse(status == 0,"no event/ censored","event")

) |>

tbl\_summary(by = status,

type = all\_dichotomous() ~ "categorical",

statistic = list(

all\_categorical() ~ "{n} ({p})%"

)) |>

add\_overall()

#Distribution of Continuous variables

grid.arrange(cvd.df |> ggplot( aes(x = age)) +

geom\_histogram(binwidth = 1, fill = "black", color = "white") +

labs(title = "Age Distribution") +

theme(plot.title = element\_text(hjust = 0.5)),

cvd.df |> ggplot( aes(x = bmi)) +

geom\_histogram(binwidth = 1, fill = "black", color = "white") +

labs(title = "BMI Distribution",

x = "Body Mass Index") +

theme(plot.title = element\_text(hjust = 0.5)) +

theme(plot.title = element\_text(hjust = 0.5)),ncol = 2, heights = c(1,

0.2, 0))

# Cox PH Modelling with RMS package ----------------------------------------------------

--------

units(cvd.df$tevent) <- "years"

dd = datadist(cvd.df)

options(datadist='dd')

surv.obj <- Surv(cvd.df$tevent,cvd.df$status)

model\_simple <- cph(surv.obj ~ age + sex + smoking\_status + bmi + diabetic + ckd +

atrial\_fibrillation,

data =cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow',

time.inc = 5)

summary(model\_simple)

# Proportional Hazard Test Simple Model ------------------------------------------------

ph.test.simple <- cox.zph(model\_simple)

ph.test.simple

# Linearity Test for age --------------------------------------------------

normal\_age\_model <- cph(surv.obj ~ age,

data = cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow')

rcs\_age\_model <- cph(surv.obj ~ rcs(age,3),

data = cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow')

lr\_test = lrtest(normal\_age\_model,rcs\_age\_model)

# Linearity Test for bmi --------------------------------------------------

normal\_bmi\_model <- cph(surv.obj ~ bmi,

data = cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow')

rcs\_bmi\_model <- cph(surv.obj ~ rcs(bmi,3)

data = cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow')

lr\_test = lrtest(normal\_bmi\_model,rcs\_bmi\_model)

# Fitting Complex Model ---------------------------------------------------

model\_complex <- cph(surv.obj ~ rcs(age,3) + strat(sex) + smoking\_status + bmi +

diabetic + ckd + atrial\_fibrillation,

data =cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow',

time.inc = 5)

# Proportional Hazard Test Simple Model -----------------------------------

ph.test.complex <- cox.zph(model\_complex)

ph.test.complex

# AIC TESTS ---------------------------------------------------------------

AIC(model\_simple)

AIC(model\_complex)

#AIC(final\_model\_complex)

# STEP WISE MODEL SELLECTION ----------------------------------------------

final <- fastbw(model\_complex, rule = "aic")

# FINAL OPTIMAL MODEL -----------------------------------------------------

final\_model\_complex <- cph(surv.obj ~ rcs(age,3) + strat(sex) + bmi + diabetic + ckd,

data = cvd.df ,

surv = TRUE,

x = TRUE,

y = TRUE,

method = 'breslow',

time.inc = 5)

adjusted\_hazard\_ratios <- exp(final\_model\_complex$coefficients)

ahr\_conf\_int <- exp(confint(final\_model\_complex))

f <- update(final\_model\_complex,x=TRUE,y=TRUE)

# INTERNAL VALIDATION WITH BOOTSTRAP METHOD -------------------------------

set.seed(123) # Set seed for reproducibility

validation\_results\_boot <- validate(final\_model\_complex,method = "boot", B = 200,u = 5)

reds <- predict(f, type = "lp", times = 5)

c\_index <- rcorr.cens(preds,surv.obj)

c\_indexx <- c\_index['C Index']

se <- c\_index['S.D.']/2

low <- c\_indexx-1.96\*se

hi <- c\_indexx +1.96\*se

# INTERNAL CALIBRATION WITH BOOTSTRAP METHOD ------------------------------

calibrate\_results\_boot <- calibrate(final\_model\_complex,method = "boot",dxy = T, B =

200,u = 5)

plot(calibrate\_results\_boot,las = 1)

# MODEL PRESENTATION WITH NOMOGRAM ----------------------------------------

nom <- nomogram(final\_model\_complex)

plot(nom)

#####Nomogram

surv <- Survival(final\_model\_complex)

surv2y <- function(x) surv(5, lp = x)

ss <- c(-2, -1,0.05, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 1,2)

nomo\_gram <- nomogram(final\_model\_complex, fun = surv2y, fun.at = ss, lp = TRUE,

funlabel = "5 year survival")

plot(nomo\_gram)

survdiff(surv.obj~sex,data = cvd.df) |> plot()

cox\_model <- survfit(surv.obj ~ sex, data = cvd.df)

ggsurvplot(cox\_model, data = cvd.df, risk.table = TRUE)

# Create a survival plot using survplot

survplot(cox\_model, data = cvd.df, col = c("blue", "red"), lty = c(1, 2), lwd = c(2, 2),

xlab = "Time", ylab = "Survival Probability", main = "Survival Curve by Sex")