UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Health inequalities in cerebro and cardiovascular health – an analysis of the e_COR study

Carolina Sofia Afonso dos Santos

Mestrado em Bioestatística

Versão Final

Trabalho de Projeto orientado por:

Orientador: Profª Doutora Marília Antunes
Coorientador: Doutora Ana Catarina Alves

2022

Resumo

Segundo a Organização Mundial de Saúde (OMS), desigualdades em saúde são diferenças sistemáticas no estado de saúde de diferentes grupos populacionais com custos significativos sociais e económicos para os indivíduos e sociedades. Estas diferenças são injustas e podem ser atenuadas através da ação de políticas governamentais, pelo que é necessário compreender qual o seu impacto.

As doenças cardiovasculares (DCV) são um conjunto de doenças que afeta o cérebro, coração e vasos sanguíneos, incluindo doença coronária, doença cérebro-cardiovascular, doença cardíaca reumática, entre outras. Mais do que quatro em cada cinco mortes causadas por DCV são devido a ataques cardíacos e acidentes vasculares cerebrais (AVC), sendo que um terço destas mortes ocorre prematuramente em pessoas com menos de 70 anos (OMS). Os eventos de interesse considerados neste estudo foram os seguintes: AVC, enfarte agudo do miocárdio e doença arterial periférica.

Em Portugal em 2019, os AVCs foram a principal causa de morte, representando 9.8% da mortalidade total, segundo o Instituto Nacional de Estatística (INE). As mortes por enfarte do miocárdio foram cerca de 3.8% da mortalidade total e estima-se que foram perdidos cerca de 10670 potenciais anos de vida devido às doenças cerebro-cardiovasculares. Para além de serem interessantes por serem muito prevalentes em toda a Europa, as doenças cérebro-cardiovasculares são impactadas pelo estilo de vida e comportamentos dos indivíduos. Vários estudos já concluíram que uma parte substancial do risco de desenvolver uma doença cérebro-cardiovascular pode ser reduzido através de escolhas pessoais e de abordagens preventivas, como adotar estilos de vida saudáveis.

Sabe-se também que a posição socioeconómica de um indivíduo tem um potencial efeito no seu estado de saúde e também nos cuidados de saúde que recebe, o que significa que as desigualdades socioeconómicas são determinantes de saúde. O Índice Europeu de Privação (EDI) é uma forma de quantificar o nível de privação dos indivíduos e, por isso, a versão adaptada à realidade portuguesa, desenvolvida em 2016 por uma equipa multinacional e multidisciplinar de investigadores, foi incorporada neste trabalho.

O objetivo principal deste trabalho é avaliar o efeito das desigualdades em saúde, nomeadamente as associadas à baixa escolaridade e aos níveis de privação na doença cérebro-cardiovascular. Para alcançar este objetivo foram utilizados os dados resultantes do estudo e_COR, realizado pelo Instituto Nacional de Saúde Doutor Ricardo Jorge, entre 2012 e 2014, no âmbito da prevenção cardiovascular e com o objetivo de estimar a prevalência dos fatores de risco cardiovascular.

No presente trabalho foi estudada a passada ocorrência de pelo menos um evento cérebro-cardiovascular e o número de fatores de risco que cada um dos participantes no estudo e_COR apresentava. As variáveis em estudo são algumas das variáveis recolhidas durante o estudo e_COR, referentes a características demográficas, físicas, metabólicas, historial médico e estilos de vida dos participantes. Também foram recolhidas variáveis que caracterizam o acesso aos cuidados de saúde, extraídas a partir do INE, e variáveis relativas à condição socioeconómica que resultam do EDI adaptado à realidade portuguesa.

Para modelar a ocorrência de um evento cérebro-cardiovascular foi utilizado o método de regressão logística uma vez que se adequa à natureza binária da variável resposta. Antes da modelação dos dados foram avaliadas possíveis correlações entre as variáveis em estudo. As variáveis que foram incluídas no Modelo Linear Generalizado foram selecionadas através do processo de Seleção Stepwise, considerando um p-value de entrada de 0.20 e um p-value de saída de 0.25. As variáveis de confundimento, sexo e idade, foram as primeiras a serem introduzidas no modelo e independentemente do seu p-value permaneceram sempre no modelo. Após o processo Stepwise foram introduzidas as variáveis relacionadas com níveis de educação e privação. Ao modelo final foi adicionado o efeito aleatório da região de residência. Uma vez que se verificou um grande desequilíbrio entre casos e não-casos de doença cérebro-cardiovascular, foi utilizada a técnica de SMOTE para criar um conjunto de

dados mais equilibrados e o mesmo processo de modelação foi aplicado. Para a validação dos modelos obtidos foi calculada a curva ROC com a respetiva AUC e também foi realizada validação cruzada.

Para modelar o número de fatores de risco foi utilizada a regressão de Poisson uma vez que é um dos métodos indicados quando a variável resposta é uma contagem. Neste modelo foram incluídas, mais uma vez, as variáveis de confundimento, sexo e idade, bem como as variáveis relativas a desigualdades, níveis de educação e de privação. Foram realizadas comparações múltiplas para avaliar de que forma as variáveis associadas às desigualdades afetam o número de fatores de risco que cada indivíduo apresenta.

Os eventos cérebro-cardiovasculares registados ocorreram no passado e os fatores de risco que os indivíduos apresentam estão no presente, pelo que a interpretação dos modelos não é a mais convencional e algumas das variáveis teriam valores diferentes se tivessem sido avaliadas antes do evento ocorrer.

Os modelos de regressão logística sugerem que na presença de dois indivíduos com características físicas e biológicas semelhantes, a influência dos seus níveis de educação ou privação não é significativa nem são bons indicadores para distinguir indivíduos que já sofreram um evento cérebro-cardiovascular daqueles que não sofreram. O efeito aleatório da região também não é estatisticamente significativo. Uma vez que o evento já ocorreu este modelo serve sobretudo para distinguir indivíduos que já sofreram o evento daqueles que não sofreram, através das suas características.

O modelo de regressão Poisson sugere que os níveis de educação de um indivíduo afetam o número de fatores de risco que o mesmo apresenta, isto é, verificou-se que indivíduos com escolaridade mais avançada apresentavam menor número de fatores de risco.

Palavras-chave: aparelho circulatório, doença cérebro-cardiovascular, regressão logística, regressão Poisson, efeitos aleatórios

Abstract

According to the World Health Organization (WHO), health inequalities are systematic differences in the health status of different population groups with significant social and economic costs for individuals and societies. These differences are unfair and can be mitigated through the action of government policies, so it is necessary to understand their impact.

Cardiovascular diseases (CVD) are a group of disorders of the brain, heart and blood vessels and include coronary heart disease, cerebro-cardiovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age (WHO). The events of interest considered in this work were stroke, acute myocardial infarction, and peripheral arterial disease.

In Portugal in 2019, strokes were the number one cause of death representing 9.8% of the total mortality and the deaths by myocardial infarction represented 3.8% of the total mortality, according to INE. It is estimated that about 10 670 potential years of life were lost due to cerebro-cardiovascular diseases in Portugal in 2019. In addition to being interesting for being very prevalent across Europe, cerebrovascular diseases are impacted by the lifestyle and behavior of individuals. Several studies have concluded that a substantial part of the risk of developing cerebro-cardiovascular disease can be reduced through personal choices and preventive approaches, such as adapting healthy lifestyles.

An individual's socioeconomic position has a potential effect on their health status and also on the health care they receive, which means that socioeconomic inequalities are determinants of health. The European Deprivation Index (EDI) is a way of quantifying the level of deprivation of individuals and, therefore, the version adapted to the Portuguese reality, developed in 2016 by a multinational and multidisciplinary team of researchers, was incorporated in this work.

The main objective of this work is to evaluate the effect of health inequalities, namely those associated with low education and levels of deprivation in cerebro-cardiovascular disease. To achieve this objective, the main set of data used resulted from the e_COR study, developed by the Instituto Nacional de Saúde Doutor Ricardo Jorge, between 2012 and 2014, within the scope of cardiovascular prevention and with the objective of estimating the prevalence of cardiovascular risk factors.

In the present work, the past occurrence of at least one cerebro-cardiovascular event and the number of risk factors that each of the participants in the e_COR study had were studied. The variables under study are some of the variables collected during the e_COR study, referring to demographic, physical, metabolic, medical history, and lifestyle characteristics of the participants. Some other variables were collected such as variables that characterize access to healthcare, extracted from INE, and variables related to socioeconomic status, that resulted from the EDI adapted to the Portuguese context.

To model the occurrence of a cerebro-cardiovascular event, the logistic regression method was used, as it suits the binary nature of the response variable. Before modeling the data, possible correlations between the variables under study were evaluated. The variables that were included in the Generalized Linear Model were selected through the Stepwise Selection process, considering an inclusion p-value of 0.20 and an exclusion p-value of 0.25. The confounding variables, sex and age, were the first to be introduced in the model and, regardless of their p-value, they always remained in the model. After the Stepwise process, variables related to levels of education and deprivation were introduced. The random effect of the region of residence was added to the final model. Since there was a large imbalance between cases and non-cases of cerebro-cardiovascular disease, the SMOTE technique was used to create a more balanced dataset and the same modeling process was applied. For the validation of the models obtained, the ROC curve with the respective AUC was calculated and cross validation was also performed.

Poisson regression was used to model the number of risk factors, since it is one of the methods indicated when the response variable is a count variable. In this model, once again, confounding variables, sex and age were included, as well as the variables related to inequalities, education levels and deprivation levels. Multiple comparisons were performed to assess how variables associated with inequalities affect the number of risk factors that each individual has.

The recorded cerebro-cardiovascular events that occurred happened in the past and the risk factors that individuals have are in the present, so the interpretation of the models is not the most conventional and some of the variables would have different values if they had been evaluated before the event occurred.

Both logistic regression models suggest that in the presence of two individuals with similar physical and biological characteristics, the influence of their levels of education or deprivation is not significant, and these are not good indicators to distinguish individuals who have already suffered a cerebro-cardiovascular event from those who have not. The random effect of the region is also not statistically significant. Since the event has already occurred, this model is useful mainly to distinguish individuals who have already suffered the event from those who have not suffered, through their characteristics.

The Poisson regression model suggests that an individual's education levels affect the number of risk factors that an individual has, i.e., it was found that individuals with more advanced education had a lower number of risk factors.

Keywords: circulatory system, cerebro-cardiovascular disease, logistic regression, Poisson regression, random effects

Index

List of Figures

List of Tables

Abbreviations and Acronyms

AIC	Akaike Information Criterion
AUC	Area Under the Curve
BMI	Body Mass Index
CCV	Cerebro-cardiovascular
CI	Confidence Interval
CVD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
EDI	European Deprivation Index
EU	European Union
EU-SILC	European Union Statistics on Income and Living Conditions
FN	False Negative
FP	False Positive
GLM	General Linear Model
HDL	High-density lipoprotein
INE	Instituto Nacional de Estatística
INSA	Instituto Nacional de Saúde Doutor Ricardo Jorge
IPAQ	International Physical Activity Questionnaire
LDL	Low-density lipoprotein
LVT	Lisbon and Vale do Tejo
OR	Odds Ratio
ROC	Receiver Operating Characteristic
SBP	Systolic Blood Pressure
SD	Standard Deviation
SMOTE	Synthetic Minority Oversampling Technique
TN	True Negative
TP	True Positive
VIF	Variance Inflation Factor
WHO	World Health Organization

Chapter 1 - Introduction

This chapter will begin by establishing the differences between inequity and inequality and how these are described in the context of health. The definition of cerebro-cardiovascular disease will also be presented as well as the events considered of interest within the cerebro-cardiovascular diseases in the context of this work. An overall description of the data used to conduct the analyses and the objectives as well as the analysis plan will be presented. The chapter ends with an overview of the entire work.

1.1 Inequity vs. Inequality

According to Global Health Europe, inequity refers to unfair and avoidable differences arising from poor governance, corruption, or cultural exclusion while inequality simply refers to the uneven distribution of health or health resources as a result of genetic or other factors or the lack of resources. Inequity can be measured in terms of the inequality of health or resources, raising the question of when does inequality in health or resources constitute inequity.[1] Health inequity can be considered as a specific type of health inequality, when the health differences are considered to be unjust.[2]

The World Health Organization (WHO) defines health inequities as systematic differences in the health status of different population groups that have significant social and economic costs both to individuals and societies. Social factors such as education, employment status, income level, gender and ethnicity have an influence on how healthy a person is. In all countries – whether low-, middle- or high-income – there are wide disparities in the health status of different social groups. The lower an individual's socio-economic position, the higher their risk of poor health. [3] Due to the fact that these differences are unfair and can be attenuated by government policies it is very important to understand them and their real impact.

1.2 Cerebro-cardiovascular diseases

Cardiovascular diseases (CVD) are a group of disorders of the brain, heart and blood vessels and include coronary heart disease, cerebro-cardiovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. [4]

Stroke, acute myocardial infarction and peripheral arterial disease were considered to be the events of interest within the cerebro-cardiovascular disease in this study. A stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients, resulting on the death of brain cells in minutes. [5] An acute myocardial infarction, commonly known as a heart attack, is myocardial necrosis resulting from acute obstruction of a coronary artery, i.e., tissue damage in the heart muscle when blood flow is interrupted, usually caused by a blockage in the coronary arteries. [6] Peripheral arterial disease is a circulatory problem where the blood flow to the limbs is reduced due to narrowed arteries. [7]

In 2016, there were 1.68 million deaths in the European Union (EU) from diseases of the circulatory system, equivalent to 37.1% of all deaths, constituting the most prevalent cause of death. [8] Diseases of the circulatory system are one of the main causes of mortality in each of the EU member states. [8]

In recent years, there has been a decrease in the proportion of deaths caused by diseases of the circulatory system in the total number of deaths, from 31.9% in 2009 to 29.9% in 2019. Still, in 2019, strokes continued to be the cause of the highest number of deaths, according to the Portuguese National

Institute of Statistics (INE), representing 9.8% of the total mortality and a rate of 106.5 resident deaths per 100 mil inhabitants. The deaths by acute myocardial infarction represented 3.8% of the total mortality and almost 60% of ischemic heart disease deaths, in the same period. Of the total deaths from cerebro-cardiovascular diseases, 93.6% were of people aged 65 and over and 82.5% of people aged 75 and over. In 2019, 10 670 potential years of life were lost due to cerebro-cardiovascular diseases. [9]

In addition to being interesting due to their high prevalence in all of Europe, the interest in studying cerebro-cardiovascular diseases also lies on the fact that lifestyle choices and behaviours have a strong impact in their development. Through the INTERSTROKE study, a standardised case-control study conducted worldwide with the objective of establishing the association of known and emerging risk factors with stroke, it was found that five risk factors collectively account for more than 80% of the world's risk. [10] These include high blood pressure, current smoking, abdominal obesity and diet, leading to the conclusion that a healthy lifestyle could reduce the burden of smoke. [10] Another important study was the INTERHEART, a standardised case-control study conducted worldwide with the objective of evaluating the effects of potentially modifiable risk factors on myocardial infarction, that found that most of the risk factors for myocardial infarction, such as high blood pressure, diabetes, and abdominal obesity, are associated with lifestyle factors. [11] These risk factors can be prevented through various approaches such as regular physical activity and healthy diets. [11]

The main conclusion of these two studies is that a substantial part of the risk of developing a cerebro-cardiovascular disease can be reduced through personal choices and through preventative approaches, such as adopting healthier habits.

1.3 Socioeconomic Conditions

As mentioned above, an individual’s socioeconomic position has a potential effect on their health status and also on the health outcomes and care they receive, meaning that socioeconomic inequalities are major health determinants. Although it is not a consensual concept, deprivation is still considered a relative concept. In 1970, economist John Townsend argued that deprivation is a relative concept that refers to the state of being in poverty relative to the community or the nation where one lives. [12]

There are few studies dealing with economic inequality in Portugal. However, the number of ecological deprivation indexes has grown significantly over the last decade. This diversity of indexes hinders study comparability and increases the difficulty in addressing the topic. The goal of the development of the European Deprivation Index (EDI) was to create an ecological deprivation index for Portugal, Spain, England, and the small areas of France. The index was developed using the methodology of the Townsend Theorization of Deprecation and it was based on the data collected by the European Union's poverty survey. [12]

The EDI project adapted to Portugal involved three key methodological step-wise stages. These included the creation of an individual deprivation indicator, the establishment of a baseline index, and the analysis of the available data. There is no precise definition for individual deprivation. Instead, the concept of fundamental needs was used to measure it. If the majority of people have these needs, then those who have not are in disadvantage. Several items were not possessed by most households in the Portuguese EU-SILC survey. Only those possessed by the majority were deemed fundamental needs. Aside from being related, deprivation is also associated with poverty. [12]

The construction of the Portuguese Deprivation Index focused on the needs that were associated with both subjective and objective poverty. The income needed to meet these fundamental needs was then determined by the household's equivalized disposable income. It was found that 20.7% of the households were considered objectively poor. The question "ability to make ends meet" was asked by the EU-SILC Likert-scale. The threshold at which people felt poor was determined by univariable logistic regressions. In Portugal, 15.7% of the households were considered subjectively poor. Only the

needs significantly associated with subjective and objective poverty were included. The second step of the project involved the use of ecological data collected from the 2001 census in Portugal. The data was collected at the upper aggregation level and, in most cases, at census tract block groups. Logistic regression was performed to determine which of the various ecological variables would be included in the Portuguese deprivation index and the score resulted from equation 1.1. [12]

$$\begin{aligned}\text{Score} = {} & \%\ \text{non-owned households (z-score)} \times 1.193 \\ & + \%\ \text{households without indoor flushing (z-score)} \times 1.456 \\ & + \%\ \text{residents with low education level} (\leq \text{6th grade}) \text{ (z-score)} \times 1.292 \\ & + \%\ \text{household with 5 rooms or less (z-score)} \times 0.404 \\ & + \%\ \text{unemployed looking for a job (z-score)} \times 0.376 \\ & + \%\ \text{female residents aged 65 years or more (z-score)} \times 0.255 \\ & + \%\ \text{households without bath/shower (z-score)} \times 0.060 + \%\ \text{residents} \\ & \text{employed in manual occupations (z-score)} \times 0.013\end{aligned} \tag{1.1}$$

The Portuguese EDI had the following distribution: minimum = -8.155; maximum = 17.249; mean = 0.000 and standard deviation = 2.283. The quintiles of the EDI score as cut-offs were the following: 1 (-8.155 to -1.774); 2 (-1.773 to -0.605); 3 (-0.605 to 0.338); 4 (0.338 to 1.581) and 5 (1.582 to 17.249). The first quintile represented the lowest level of deprivation and included 20.9% of the national population, the second included 21.2%, the third included 20.5%, the fourth 19.5% and the fifth and most deprived included 18% of the population (Figure 1.1). [12]

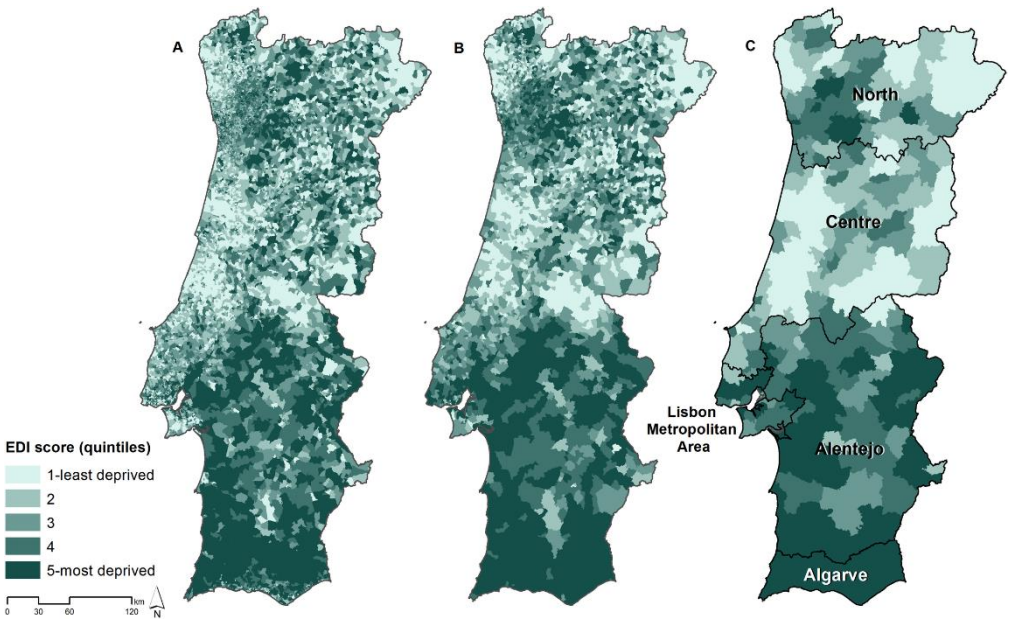

Figure 1.1: Spatial distribution of the European Deprivation Index for Portuguese small-areas in Continental Portugal. (A: Census block groups; B: Parishes; C: Municipalities).[12]

1.4 e_COR, Sociodemographic and Healthcare Access Data

The main data used in this work is from a study of the prevalence of cardiovascular risk factors of the Portuguese population, called e_COR, conducted by Instituto Nacional de Saúde Ricardo Jorge. The main aspects of this study will be detailed in Chapter Two. In addition to this, sociodemographic and access to healthcare data were collected at the geographical scale of the municipality, provided by the Instituto Nacional de Estatística (INE), namely:

- Number of healthcare centres
- Number of doctors in healthcare centres
- Number of nurses in healthcare centres
- Resident population

1.5 Objective and Analysis Plan

The main objective of this project is to evaluate the effect of health inequalities, namely those associated with low education and levels of deprivation on cerebro-cardiovascular disease.

Regarding cerebro-cardiovascular disease, the number of risk factors that the individuals present will be studied, as well as the past occurrence of at least one cerebro-cardiovascular event at the time of the survey, using general linear models. If it proves to be more suitable, the Generalized Additive Models class will also be considered.

Linear regression analysis is one of the main statistical modeling techniques and its main objective is to analyze the existence of a linear relationship between a response variable and one or more explanatory variables. It makes it possible to understand the causes of variation in a phenomenon and predict its behavior according to explanatory variables. This analysis requires strong assumptions, such as normality, independence and homoscedasticity of errors, and non-compliance with them is quite common and that is why the Generalized Linear Models emerged, which are an extension of the regression models. The distribution family to use in GLM depends on the type of values that the variable can take, and in this case, we have: the number of risk factors of individuals, this being count data and, therefore, the Poisson regression model will be used; the occurrence of at least one cerebro-cardiovascular event at the time of the survey, this being binary data (presence/absence) and therefore the logistic regression model will be applied. In these models, the variables that allow the characterization of the access to health will be considered as independent variables. The control of confounding effects such as sex and age will be considered, among others, to be evaluated.

It is expected to find differences between the sampled due to their different levels of deprivation and social differences, namely academic qualifications - more developed regions usually have more university graduates and less developed regions usually have lower levels of education.

1.6 Overview

This work is divided in five chapters. Chapter 2 is dedicated to describing the data used, from the e_COR study, from INE and from the European Deprivation Index adapted to Portugal.

Chapter 3 details all the statistical analyses conducted to analyse the data, the variables of interest and how they were modelled.

Chapter 4 contains the results of the statistical modelling and chapter 5 is dedicated to the discussion and conclusions.

Chapter 2 - The Data

This chapter is dedicated to describing the e_COR study, its scope, objectives, and main results. It also presents a brief description of the socioeconomic, sociodemographic and access to healthcare data also used to conduct the analyses.

2.1 e_COR study

As mentioned previously in chapter 1, cerebro-cardiovascular diseases have a high prevalence in the Portuguese population, with stroke being the number one cause of death, representing 9.8% of the total mortality, and acute myocardial infarction representing 3.8% of the total mortality. [9] Taking this into consideration, between 2012 and 2014, the e_COR study was developed by INSA, in the context of cardiovascular prevention. This study also aimed at estimating the prevalence of major cardiovascular risk factors as well as determining the overall cardiovascular risk. This was an observational and cross-sectional epidemiological study, whose non-stratified sample was intended to be representative of the Portuguese population, being characterized by sex and age group. The study included 1688 participants, 848 men and 840 women, from five regions within Mainland Portugal. When the population was sampled, there was an effort to have roughly the same number of individuals in each region per age group and gender as can be observed in Tables 2.1 and 2.2.

Table 2.1: Study participants by region and sex

Region	Men	Women	Total
North	169	171	340
Center	170	168	338
Lisboa e Vale do Tejo	184	165	349
Alentejo	166	177	343
Algarve	159	159	318
Total	848	840	1688

Table 2.2: Study participants by region and age group

Region	18-34	35-64	65-79	Total
North	107	118	115	340
Center	88	132	118	338
Lisboa e Vale do Tejo	102	126	121	349
Alentejo	113	111	119	343
Algarve	93	119	106	318
Total	503	606	579	1688

The data was collected through a questionnaire, and physical and clinical exams in eighteen healthcare centers of fourteen Portuguese municipalities, as shown in Table 2.3

The main goal of the e_COR study was to estimate the prevalence of the main risk factors for cerebro-cardiovascular disease in order to encourage cardiovascular prevention, by the individuals and government prevention policies, and assess people's perception of their state of health and/or disease, treatment and control of the following pathologies: diabetes mellitus, hypercholesterolemia, hypertriglyceridemia and arterial hypertension.

Table 2.3: Healthcare Centers, Municipalities and Regions

Region	Municipality	Healthcare Center
North	Porto	São João do Porto
		Aldoar
	Maia	Maia/Águas Santas
		Castelo da Maia
Center	Pombal	Pombal
	Montemor-o-Velho	Montemor-o-Velho
	Soure	Soure
Lisboa and Vale do Tejo	Odivelas	Odivelas
	Lisboa	Olivais
		Graça
Alentejo	Montemor-o-Novo	Montemor-o-Novo
	Évora	Center of Évora
	Ponte de Sor	Ponte de Sor
		Montargil
Algarve	Faro	Faro – Polo I
	Olhão	Olhão
	Silves	Silves
	Portimão	Portimão

All the analyses in this study were conducted using SPSS and considering a significance level of 5%. The choice of variables to include in the study was based on the already known risk factors to cerebro-cardiovascular disease – biological and non-biological. This study was particularly important since there was no study yet, in Portugal, that had assessed the prevalence of the main cerebro-cardiovascular risk factors in a population sample of this size. The collection of information was very extensive, resulting in a database with around six-hundred variables of different kinds:

- Social characteristics, such as academic qualifications, professional activity, marital status
- Lifestyle-associated, such as smoking habits, alcohol consumption, diet, and physical activity levels
- Metabolic parameters, such as total-cholesterol, HDL-c, LDL-c, triglycerides, glucose
- Hematological parameters, such as leukocytes, neutrophils, lymphocytes, monocytes, platelets
- Physiological parameters, such as heart rate, systolic and diastolic blood pressure
- Physical characteristics, such as weight, height, body mass index, waist perimeter
- Medical history, such as present and past diseases, prescribed medication, and history of cerebro-cardiovascular diseases

The metabolic, hematologic and physiological parameters as well as physical characteristics were collected in an interview setting and not self-reported.

According to the e_COR study´s report the risk factors for cerebro-cardiovascular diseases are the following:

- Obesity, when the body mass index is above 30 kg/m^2
- Abdominal obesity, when the waist perimeter is above 108 cm for men and 88 cm for women
- Hypertension, when SBP is above 140 mmHG or DBP is above to 90 mmHG
- Total cholesterol, above 240 mg/dL and above 190 mg/dL
- LDL-cholesterol, above 160 mg/dL and above 115 mg/dL
- HDL-cholesterol, below 40 mg/dL

- Triglycerides, above 200 mg/dL and above 150 mg/dL
- Diabetes, glucose levels above 126 mg/dL
- Physical activity, when at a low level (defined according to IPAQ classification)
- Smoking habits, when an individual is a regular or occasional smoker
- Inadequate diet, when an individual does not consume 5 or more fruit and vegetable pieces per day
- High consumption of alcohol, when it is above 10 gr for women and 20 gr for men
- History of cerebro-cardiovascular diseases, when an individual has first degree relatives (parents, siblings and/or children) presenting premature cerebro-cardiovascular diseases.

In addition to this classification, it was also considered that an individual had the risk factor if they were taking medication for it, for example, if an individual had normal total cholesterol levels but was taking medication to control those levels, it was considered that this individual had the risk factor related to high total cholesterol, that increases the chances of developing cerebro-cardiovascular diseases.

At the end of the e_COR study it was reported that 68% of the Portuguese population presented two or more risk factors and 22% presented four or more. Overall, it was found that there is a high degree of ignorance regarding own pathologies, prescribed medication and a low control rate of the risk factors contributing to the development of cerebro-cardiovascular diseases, thus the need to define policies and develop campaigns to improve health literacy.

Table 2.4: Summary table of the prevalences of the risk factors [13]

Risk Factors for CVD	Estimated Prevalence	C.I. 95%	Sample Error
Pre-obesity/Obesity	62.1%	59.8 - 64.4	2.3
Hypertension	43.1%	40.7 - 45.5	2.4
Dyslipidemia: LDL-cholesterol ≥ 160 mg/dL	31.5%	29.3 – 33.6	2.2
Dyslipidemia: LDL-cholesterol ≥ 115 mg/dL	55.0%	53.0 – 57.8	2.1
Low physical activity level	29.2%	27.0 – 31.4	2.2
Smoking habits	25.4%	23.3 – 27.3	2.1
Dyslipidemia: HDL-cholesterol< 40 mg/dL	14.0%	12.3 – 15.7	1.7
History of cerebro-cardiovascular diseases	11.8%	10.3 – 13.3	1.5
Diabetes	8.9%	7.5 – 10.3	1.4
Dyslipidemia: Triglycerides ≥ 200 mg/dL	8.6%	7.3 – 9.9	1.3
Dyslipidemia: Triglycerides ≥ 150 mg/dL	18.6%	16.3 – 20.9	1.2

Even though the data from this study has already been analyzed under different scopes, it has never been used to evaluate the impact of healthcare access inequalities on the past occurrence of cerebro-cardiovascular diseases and on the number of risk factors an individual presents, hence the interest of this work.

The variables that represent each one of these risk factors were already present in the e_COR database, being dichotomous variables that take the value 0 when an individual does not have the risk factor and take the value 1 when an individual has the risk factor. All the risk factors were summed for each individual, creating a new variable that represented the total number of risk factors that each participant presents.

The past occurrence of a cerebro-cardiovascular event was defined as an individual having had at least one of the following: stroke, peripheral arterial disease, or acute myocardial infarction. This new variable, **cvevent**, is also dichotomous, assuming the value 0 when an individual did not have a previous

cerebro-cardiovascular event and assuming the value 1 when an individual had a previous cerebro-cardiovascular event. This was one of the variables considered to be of interest and it was modeled through logistic regression.

Another variable was created, that represented the total number of risk factors an individual presented, **nrfactors**. This was other variable of interest, and it was modeled through Poisson regression.

2.2 Socioeconomic, Sociodemographic and Access to Healthcare Data

Some variables related to healthcare access were collected as well, from INE, being the following:

- Number of healthcare centers per municipality, as an indicator of accessibility
- Number of doctors in healthcare centers per 10 000 inhabitants, as an indicator of availability of medical care
- Number of nurses in healthcare centers per 10 000 inhabitants, as an indicator of nursing care availability
- Resident population per municipality, as an indicator of size and development of the municipality

The year of collection of the above-mentioned data was 2012, the year the e_COR study began. Two extra variables regarding socioeconomic position were also included in the study, and these were the Portuguese EDI score and quintile, per municipality. The Portuguese EDI was developed in 2016 by a multinational and multidisciplinary team. The data from the Portuguese EDI can be easily accessed online (https://figshare.com/s/3a4226d520df3b18cb71). These variables and all the other ones used to conduct the analyses are listed in Table 2.5.

Table 2.5: Variables included in the dataset used to conduct the analyses

Variable	Description	Type	Codification
cvevent	Occurrence of at least one past cerebro-cardiovascular event	Categorical	Yes, No
sex	Sex of the participant	Categorical	Male Female
hypertensorsmed	Medication for hypertension	Categorical	Yes, No
cholmed	Medication for cholesterol	Categorical	Yes, No
trigmed	Medication for triglycerides	Categorical	Yes, No
diabetesmed	Medication for diabetes	Categorical	Yes, No
waistrf	Risk factor associated with waist perimeter superior to 108 cm for men and 88 cm for women	Categorical	Yes, No
bmirf	Risk factor associated body mass index superior to 30 kg/m^2	Categorical	Yes, No
hypertensionrf	Risk factor associated with SBP superior 140 mmHG or DBP superior to 90 mmHG or taking medication for hypertension	Categorical	Yes, No
hdlrf	Risk factors associated with HDL-cholesterol < 40 mg/dL or taking medication for cholesterol	Categorical	Yes, No
ldl160	Risk factor associated with LDL-cholesterol > 160 mg/dL or taking medication for cholesterol	Categorical	Yes, No
hct240	Risk factor associated with total cholesterol > 240 mg/dL or taking medication for cholesterol	Categorical	Yes, No

htg200	Risk factor associated with triglycerides > 200 mg/dL or taking medication for triglycerides	Categorical	Yes, No
diabetesrf	Risk factor associated with glucose > 126 mg/dL or taking medication for diabetes	Categorical	Yes, No
parf	Risk factor associated with low levels of physical activity	Categorical	Yes, No
dietrf	Risk factor associated with having an inadequate diet	Categorical	Yes, No
historyrf	Risk factor related to family history of cerebro-cardiovascular disease	Categorical	Yes, No
smokenr	Smoking habits risk factor	Categorical	Yes, No
alcoholrf	Risk factor related to alcohol consumption	Categorical	Yes, No
alcoholpf	Protective factor related to alcohol consumption	Categorical	Yes, No
palevel	Levels of physical activity according to IPAQ	Categorical	High, Moderate, Low
region	Region of the participant	Categorical	North, Center, Lisboa and Vale do Tejo, Alentejo, Algarve
education	Academic qualifications	Categorical	No qualifications or pre-High School, High School, University Education
age	Age of the participant (in years)	Numerical	
tchol	Total cholesterol levels (in mg/dL)	Numerical	
hdl	HDL-cholesterol levels (in mg/dL)	Numerical	
ldl	LDL-cholesterol levels (in mg/dL)	Numerical	
tg	Triglyceride levels (in mg/dL)	Numerical	
glucose	Glucose levels (in mg/dL)	Numerical	
weight	Weight (in Kg)	Numerical	
height	Height (in cm)	Numerical	
bmi	Body mass index (in kg/m^2)	Numerical	
waist	Waist perimeter (in cm)	Numerical	
sbp	Systolic blood pressure (in mmHg)	Numerical	
dbp	Diastolic blood pressure (in mmHg)	Numerical	
mealsday	Number of meals (per day)	Numerical	
wineunits	Number of wine units consumed (per day)	Numerical	
beerunits	Number of beer units consumed (per day)	Numerical	
whiteunits	Number of white drinks units consumed (per day)	Numerical	
vpa	Vigorous physical activity (days per week)	Numerical	
mpa	Moderate physical activity (days per week)	Numerical	
ediscore	Score of the EDI	Numerical	
ediquintile	Quintile of the EDI	Numerical	
nrfactors	Number of risk factors presented by the participants	Numerical	

nrhcc	Number of healthcare centers	Numerical	
nrdoctors	Number of doctors per 10 000 habitants	Numerical	
nrnurses	Number of nurses per 10 000 habitants	Numerical	
respop	Resident population	Numerical	

Chapter 3 - Methodology

This chapter is dedicated to presenting the results from the statistical analyses conducted. The two response variables of interest were mentioned in the previous chapter, **cvevent** and **nrfactors**, the first one being a dichotomous variable and the later one being a count variable. Given the nature of these variables, the most appropriate approach was to perform, respectively, Logistic Regression and Poisson Regression. Due to the imbalanced nature of the data set, Weighted Logistic Regression was conducted as well, using the SMOTE Technique to obtain a balanced data set.

3.1 Pre-selection of variables

The e_COR study database is very rich in terms of the number of variables collected, totaling around six hundred variables, so there was a need to go through all of these and select the most relevant ones to this study. These include variables related to metabolic parameters, lifestyle habits, physical characteristics, history of cardiovascular diseases and finally variables related to sociodemographic and socioeconomic variables. Even after this preliminary selection there was still a vast number of variables that could be related to one another, so, the next step was to reduce this set of predictors, identifying possible correlations and multicollinearity.

The correlations were assessed using the Pearson correlation coefficient, which is a measure of linear correlation, being the ratio between the covariance of two variables and the product of their standard deviations, ranging between -1 and 1:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov\,(X, Y)}{\sqrt{var(X) \cdot var(Y)}} \tag{3.1}$$

where $\bar{x}$ and $\bar{y}$ represent the arithmetic mean of the variables *x* and *y*, respectively. [14]

The degree of correlation can be interpreted as perfect, if the coefficient absolute value is near 1; high degree if the coefficient absolute value is between 0.50 and 1; moderate degree if the coefficient absolute value is between 0.30 and 0.49; low degree if the coefficient absolute value is inferior to 0.29 and no correlation when the value is zero. [15]

Multicollinearity exists when an independent variable presents high correlation with one or more of the other independent variables present in a model and this constitutes an issue because it diminishes the statistical significance of the variable. Multicollinearity was evaluated using the Variance Inflation Factor (VIF), that is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone and it quantifies the severity of multicollinearity. The VIF can be computed for each predictor in a predictive model, such as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3.2}$$

where R_i^2 is the coefficient of determination. [16]

A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.[16]

3.2 Logistic Regression

Logistic regression is the conventional adopted statistical approach when the response variable is binary, being a type of descriptive and also predictive analysis. It is used to describe data and explain the relationship between the response variable and a set of independent variables, called explanatory or predictor variables. This type of regression can also be seen as a classification algorithm to find the probability of success or failure of an event. [17] [18]

Like every method, it has its advantages and disadvantages. One of the biggest advantages is the fact that it is easy to implement and interpret. When the response variable is binary, linear regression cannot be applied due to its assumption of linearity between the response and independent variables. This issue is easily solved with the introduction of a link function. [17] [18] The most popular link function is the logit (logistic inverse transformation) function, associated with the logistic distribution, where p is the probability of the outcome.

$$logit(p) = ln\left(\frac{p}{1-p}\right), p \in (0,1) \tag{3.3}$$

The logistic regression method assumes that: the outcome is a binary or dichotomous variable, there is a linear relationship between the logit of the outcome and each predictor variables, there are no influential values (extreme values or outliers) in the continuous predictors, there are no high intercorrelations (i.e., multicollinearity) among the predictors. The general expression for a logistic regression model is the following:

$$g\big(E(Y)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_k \tag{3.4}$$

where β_0 is the *y*-intercept, the log-odds of the event *Y=1* when all the predictor variables assume the value 0 and $\beta_n x_k$ is the regression coefficient multiplied by the value of the predictor.

3.2.1 Odds and Odds Ratio

In order to interpret correctly the results from the logistic regression it is useful to understand the concepts of odds and odds ratio. Odds are defined as the ratio of the probability of success and the probability of failure (equation 3.5), and the odds ratio (OR) represent the constant effect of a predictor on the likelihood of the occurrence of an event (equation 3.6).

$$Odds = \frac{p}{1-p} \tag{3.5}$$

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{3.6}$$

There is a direct relationship between the coefficients produced by the logit function and the odds ratio because the logit can also be called log-odds since it is equal to the logarithm of the odds. The logistic regression coefficients correspond to the estimated increase in the log-odds of the response variable per one unit increase in the value of the predictor variables when these are quantitative variables. In the case of the predictor being binary, the odds ratio compares the odds of the event

occurring at two different levels of the predictor, so the exponential function of its regression coefficient is the OR associated to the presence of the variable in the outcome against its absence.

For example, in the case of having a logistic regression model with only one predictor, the expression of the model would be the one in equation (3.7) and so the coefficient of the predictor would be equal to the odds ratio, equation (3.8).

$$g\big(E(Y)\big) = \beta_0 + \beta_1 x_1 \tag{3.7}$$

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{\exp(\beta_0) \times \exp(\beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \tag{3.8}$$

3.2.2 Stepwise Selection

Stepwise Selection is a step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model, in order to select the predictors that form the best logistic model. This method incorporates the forward selection and backwards elimination procedures.

The first step is to determine a p-value for inclusion (P_i) and a p-value for exclusion (P_e) of variables in the model. A common approach is to establish $P_i = 0.20$ and $P_e = 0.25$. This way, for a variable to be considered important it does not necessarily need to be statistically significant. [19]

The response variable, Y, is regressed on each one of the predictor variables (X_i) and the X_i with the smallest p-value (inferior to the entry p-value) is added to the model. Once again, Y is regressed on the rest of the predictor variables and interesting interactions between them, and the one with the smallest p-value is added to the model. If any of the variables in the model has a p-value above the exclusion p-value, that variable should be removed. This process is repeated until there are no predictors left with a p-value inferior to the entry p-value. [19]

There are situations where certain variables will remain in the model whether they are statistically significant or not, and in these cases the stepwise selection process begins with those variables already included in the model. In this particular case, the model began with two predictors, sex and age, that should always be kept in the model, no matter their p-value. These predictors need to be included in the model due to having major influence in the incidence and outcome of diseases, especially in cerebro-cardiovascular diseases.

After each model is computed, there is a method to determine the quality of the model called Akaike Information Criterion (AIC). The AIC can be determined through the following formula:

$$AIC = 2k - 2\ln(\hat{L}) \tag{3.9}$$

where k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model. The smaller the AIC the better fitted the model is to the data. [20]

3.2.3 ROC Curve

A receiver operating characteristic curve (ROC) is a method used to evaluate the performance of binary classification algorithms, such as logistic regression. In this type of classification there can be four possible outcomes: true positive, false positive, true negative and false negative. The ROC curve is built by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. [21] The true positive rate is also designated by sensitivity (equation 3.9),

representing the probability that an actual positive will be classified as positive, and the false positive rate is also known as one minus specificity (equation 3.10):

$$True\ Positive\ Rate = Sensitivity = \frac{TP}{TP + FN} \tag{3.10}$$

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{FP + TN} \tag{3.11}$$

where *TP* is the number of true positives, *FN* is the number of false negatives, *FP* is the number of false positives and *TN* is the number of true negatives.

The information a ROC curve provides can be summarized in another single metric, the area under the curve, AUC. Usually, the higher the AUC score the better the classifier performed. An AUC of 0.5 suggests the model has no discrimination capacity to distinguish between cases and non-cases, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding. [21]

3.2.4 Cross Validation

Cross validation, also known as out-of-sample testing, is a model validation technique to assess how the results of a statistical analysis will generalize to an independent dataset. There are different types of cross validation, such as: k-fold cross validation, stratified k-fold cross validation and leave-one-out cross validation. Both k-fold and leave-one-out methods were conducted in this work.

Leave-one-out cross validation is a particular case of cross validation, where the number of folds equals the number of instances in the dataset, meaning that the algorithm will be applied once for each instance, using all others as a training set, and using the selected instance as the test set. [22]

K-fold cross validation is a method of cross validation that randomly divides the data set in k-subsets, reserves one and trains the model on all other subsets. Then, the model is tested on the reserved subset and the prediction error is recorded, repeating this process until each of the k subsets has been used as the test set. The average of the k recorded errors is computed and called the cross-validation error. K-fold cross validation is typically performed using k = 5 or k = 10. [23]

When performing these cross validation approaches, the default metrics considered to evaluate the performance of binary classification models are Accuracy and Cohen's Kappa. Accuracy corresponds to the percentage of correctly classified instances out of all instances and can be obtained through the following formula:

$$Accuracy = \frac{TP + TN|}{TP + TN + FP + FN} \tag{3.12}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives and *FN* is the number of false negatives. [24]

Cohen's kappa is used to assess the agreement between two raters, interrater reliability, and can be used to assess the performance of a classification model. [25] The interpretation of Cohen's Kappa can be seen in Table 3.1. and its calculation can be performed using the following formula:

$$K = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) \times (TP + FN) \times (FN + TN)} \tag{3.13}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives and *FN* is the number of false negatives. [25] Cohen's Kappa can be inadequate when an imbalance distribution of classes is involved. [26]

Table 3.1: Interpretation of Cohen's Kappa [25]

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-0.20	None	0-4%
0.21-0.39	Minimal	4-15%
0.40-0.59	Weak	15-35%
0.60-0.79	Moderate	35-63%
0.80-0.90	Strong	64-81%
Above 0.90	Almost Perfect	82-100%

3.3 Weighted Logistic Regression

When the distribution of labels in a data set is skewed or biased, we are in the presence of an imbalanced dataset. These datasets usually have two subgroups, the majority and the minority classes, and the distribution between the two can be slightly or highly imbalanced. When this is the case, logistic regression can still be used but it is better to perform weighted logistic regression instead, using an oversampling technique like the Synthetic Minority Oversampling Technique (SMOTE). [27]

SMOTE is an algorithm that helps in overcoming the overfitting problem posed by random oversampling, it is a technique that allows the increase of the number of cases in the minority class in a balanced way. It takes the entire dataset, but it increases the percentage of only the minority cases. [28] SMOTE synthesizes new minority instances from already existing ones. These synthetic instances are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.

In the e_COR study only 84 individuals of 1688 had a past cerebro-cardiovascular event, constituting the minority class. A balanced data set is one that has no minority nor majority class, in this case, it would be one where roughly half of the individuals had a past event of cerebro-cardiovascular disease and the other half did not. Having a balanced dataset helps to prevent skewness of the model.

If we generate four new cases per each one case already present in the dataset we will obtain five times the number of cases, that is, 420 cases of past cerebro-cardiovascular disease. To obtain an equivalent number of non-cases, we need to generate 1.25 cases from the majority class per each new case generated from the minority class (there were 336 new cases generated and 1.25 times 336 is equal to 420). There was a need to generate non-cases because of the function in R used to conduct this technique. This way, we obtain a new balanced data set with 420 individuals in each class.

3.4 Poisson Regression

In statistical modelling, Poisson regression is a generalized linear model form of regression analysis used to model count data. Poisson regression assumes that the response variable has a Poisson distribution, and that the logarithm of its expected value can be described by a linear combination of unknown parameters. A Poisson regression model is also known as a log-linear model.

$$E[Y_i|x_i] = exp(\beta^t x_i) \tag{3.14}$$

3.4.1 Multiple Comparisons

When using a Poisson regression model and there is a need to perform *post hoc* pairwise comparisons with multiple factors and more than two levels in each factor one approach is to obtain the estimated marginal means. Marginal means are the average scores from a group or subgroup. Tukey HSD (Honestly Significant Difference) is a statistical test that can be used to find means that are significantly different from each other. So, the test hypotheses are:

$$H_0: \mu_i = \mu_j, \forall i, j \quad vs \quad H_1: \exists i, j: \mu_i \neq \mu_j$$

with the following test statistic

$$\frac{Rw}{Sw} = \frac{max_i(\overline{Y_i}. - \mu_i) - min_j(\overline{Y_j}. - \mu_j)}{\sqrt{\frac{MSE}{n_c}}} \cap Tukey_{(k,n-k)} \ under \ H_0 \tag{3.15}$$

Chapter 4 – Application

This chapter begins with the results from the exploratory analysis of the data from the e_COR study as well as the data related to healthcare access from INE and deprivation level from the Portuguese EDI. It also shows the results from the stepwise selection process for the original data set and the balanced one, obtained through the SMOTE technique, as well as the results from Poisson regression.

4.1 Exploratory Analysis

The categorical variables used in this work were characterized by their absolute and relative frequencies, by region and overall (Table 4.1).

Table 4.1: Summary of the categorical variables by region and overall

Variables [a]	North (N=340)	Center (N=338)	LVT [b] (N=349)	Alentejo (N=343)	Algarve (N=318)	Overall (N=1688)
sex						
Male	169 (49.7%)	170 (50.3%)	184 (52.7%)	166 (48.4%)	159 (50.0%)	848 (50.2%)
Female	171 (50.3%)	168 (49.7%)	165 (47.3%)	177 (51.6%)	159 (50.0%)	840 (49.8%)
cvevent						
No	325 (95.6%)	323 (95.6%)	330 (94.6%)	331 (96.5%)	295 (92.8%)	1604 (95.0%)
Yes	15.0 (4.4%)	15.0 (4.4%)	19.0 (5.4%)	12.0 (3.5%)	23.0 (7.2%)	84.0 (5.0%)
alcoholrf						
No	285 (83.8%)	241 (71.3%)	276 (79.1%)	296 (86.3%)	263 (82.7%)	1361 (80.6%)
Yes	55.0 (16.2%)	97.0 (28.7%)	73.0 (20.9%)	47.0 (13.7%)	55.0 (17.3%)	327 (19.4%)
alcoholpf						
No	55.0 (16.2%)	97.0 (28.7%)	73.0 (20.9%)	47.0 (13.7%)	55.0 (17.3%)	327 (19.4%)
Yes	285 (83.8%)	241 (71.3%)	276 (79.1%)	296 (86.3%)	263 (82.7%)	1361 (80.6%)
historyrf						
No	303 (89.1%)	308 (91.1%)	307 (88.0%)	308 (89.8%)	284 (89.3%)	1510 (89.5%)
Yes	37.0 (10.9%)	30.0 (8.9%)	42.0 (12.0%)	35.0 (10.2%)	34.0 (10.7%)	178 (10.5%)
smokingrf						
No	258 (75.9%)	279 (82.5%)	262 (75.1%)	266 (77.6%)	248 (78.0%)	1313 (77.8%)
Yes	82.0 (24.1%)	59.0 (17.5%)	87.0 (24.9%)	77.0 (22.4%)	70.0 (22.0%)	375 (22.2%)
dietrf						
No	85.0 (25.0%)	125 (37.0%)	107 (30.7%)	84.0 (24.5%)	88.0 (27.7%)	489 (29.0%)
Yes	255 (75.0%)	213 (63.0%)	242 (69.3%)	259 (75.5%)	230 (72.3%)	1199 (71.0%)
parf						
No	227 (66.8%)	244 (72.2%)	252 (72.2%)	229 (66.8%)	240 (75.5%)	1192 (70.6%)
Yes	113 (33.2%)	94.0 (27.8%)	97.0 (27.8%)	114 (33.2%)	78.0 (24.5%)	496 (29.4%)
diabetesrf						
No	302 (88.8%)	295 (87.3%)	315 (90.3%)	292 (85.1%)	286 (89.9%)	1490 (88.3%)
Yes	38.0 (11.2%)	43.0 (12.7%)	34.0 (9.7%)	51.0 (14.9%)	32.0 (10.1%)	198 (11.7%)
hct240						
No	220 (64.7%)	219 (64.8%)	221 (63.3%)	216 (63.0%)	207 (65.1%)	1083 (64.2%)
Yes	120 (35.3%)	119 (35.2%)	128 (36.7%)	127 (37.0%)	111 (34.9%)	605 (35.8%)
hdlrf						
No	297 (87.4%)	302 (89.3%)	285 (81.7%)	302 (88.0%)	285 (89.6%)	1471 (87.1%)
Yes	43.0 (12.6%)	36.0 (10.7%)	64.0 (18.3%)	41.0 (12.0%)	33.0 (10.4%)	217 (12.9%)
ldl160						
No	226 (66.5%)	210 (62.1%)	218 (62.5%)	221 (64.4%)	208 (65.4%)	1083 (64.2%)
Yes	114 (33.5%)	128 (37.9%)	131 (37.5%)	122 (35.6%)	110 (34.6%)	605 (35.8%)
htg200						
No	305 (89.7%)	311 (92.0%)	318 (91.1%)	306 (89.2%)	295 (92.8%)	1535 (90.9%)
Yes	35.0 (10.3%)	27.0 (8.0%)	31.0 (8.9%)	37.0 (10.8%)	23.0 (7.2%)	153 (9.1%)
hypertensionrf						
No	184 (54.1%)	154 (45.6%)	176 (50.4%)	167 (48.7%)	188 (59.1%)	869 (51.5%)
Yes	156 (45.9%)	184 (54.4%)	173 (49.6%)	176 (51.3%)	130 (40.9%)	819 (48.5%)
bmirf						
No	132 (38.8%)	104 (30.8%)	125 (35.8%)	127 (37.0%)	129 (40.6%)	617 (36.6%)
Yes	208 (61.2%)	234 (69.2%)	224 (64.2%)	216 (63.0%)	189 (59.4%)	1071 (63.4%)
waistrf						
No	169 (49.7%)	163 (48.2%)	212 (60.7%)	205 (59.8%)	197 (61.9%)	946 (56.0%)
Yes	171 (50.3%)	175 (51.8%)	137 (39.3%)	138 (40.2%)	121 (38.1%)	742 (44.0%)
hypertensorsmed						
Yes	118 (34.7%)	124 (36.7%)	140 (40.1%)	130 (37.9%)	91.0 (28.6%)	603 (35.7%)
No	222 (65.3%)	214 (63.3%)	209 (59.9%)	213 (62.1%)	227 (71.4%)	1085 (64.3%)
cholmed						
Yes	100 (29.4%)	83.0 (24.6%)	97.0 (27.8%)	100 (29.2%)	74.0 (23.3%)	454 (26.9%)
No	240 (70.6%)	255 (75.4%)	252 (72.2%)	243 (70.8%)	244 (76.7%)	1234 (73.1%)
trigmed						

Yes	11.0 (3.2%)	5.00 (1.5%)	9.00 (2.6%)	12.0 (3.5%)	8.00 (2.5%)	45.0 (2.7%)
No	329 (96.8%)	333 (98.5%)	340 (97.4%)	331 (96.5%)	310 (97.5%)	1643 (97.3%)
diabetesmed						
Yes	35.0 (10.3%)	33.0 (9.8%)	29.0 (8.3%)	41.0 (12.0%)	23.0 (7.2%)	161 (9.5%)
No	305 (89.7%)	305 (90.2%)	320 (91.7%)	302 (88.0%)	295 (92.8%)	1527 (90.5%)
education						
None/Pre High School	161 (47.4%)	224 (66.3%)	183 (52.4%)	164 (47.8%)	163 (51.3%)	895 (53.0%)
High School	97.0 (28.5%)	61.0 (18.0%)	83.0 (23.8%)	76.0 (22.2%)	81.0 (25.5%)	398 (23.6%)
University Education	82.0 (24.1%)	53.0 (15.7%)	83.0 (23.8%)	99.0 (28.9%)	69.0 (21.7%)	386 (22.9%)
Missing	0 (0%)	0 (0%)	0 (0%)	4.00 (1.2%)	5.00 (1.6%)	9.00 (0.5%)

[a] Refer to Table 2.5 for the description of each variable
[b] Lisbon and Vale do Tejo

As can be seen in Table 4.1, the number of individuals of each sex was quite balanced overall and within each region. The majority of the individuals in all regions did not have a cerebro-cardiovascular event in the past. Regarding the risk factors, the ones associated with unbalanced diet (**dietrf**) and high body mass index (**bmirf**) were present in most of the individuals overall and within each region. When it comes to prescribed medication, most of the individuals overall and within each region appear to not be taking any medication for hypertension, cholesterol, triglycerides, or diabetes, which makes sense since these risk factors also appear to not be present in the overall sampled population.

Overall, only 84 individuals that participated in the study (5%) had a cerebro-cardiovascular event in the past. Of these 84 individuals, 52 were men and 32 were women and in both sexes the majority that had an event belonged to the older age group (65-79), as shown in Table 4.2.

Table 4.2: Occurrence of a past cerebro-cardiovascular event per sex and age group

Variable[a]	**Male**			**Female**		
	18-34 (N=238)	**35-64 (N=298)**	**65-79 (N=312)**	**18-34 (N=265)**	**35-64 (N=308)**	**65-79 (N=267)**
cvevent						
No	238 (100%)	289 (97.0%)	269 (86.2%)	265 (100%)	294 (95.5%)	249 (93.3%)
Yes	0 (0%)	9.00 (3.0%)	43.0 (13.8%)	0 (0%)	14.0 (4.5%)	18.0 (6.7%)

[a] Refer to Table 2.5 for the description of the variable

Regarding the medication, the majority of the individuals were not medicated but the ones who were belonged to the older age group and were mainly taking medication for cholesterol and hypertension, as presented in Table 4.3.

Table 4.3: Medication taken by the individuals per sex and age group

Variables [a]	**Male**			**Female**		
	18-34 (N=238)	**35-64 (N=298)**	**65-79 (N=312)**	**18-34 (N=265)**	**35-64 (N=308)**	**65-79 (N=267)**
hypertensorsmed						
Yes	5.00 (2.1%)	75.0 (25.2%)	217 (69.6%)	13.0 (4.9%)	93.0 (30.2%)	200 (74.9%)
No	233 (97.9%)	223 (74.8%)	95.0 (30.4%)	252 (95.1%)	215 (69.8%)	67.0 (25.1%)
cholmed						
Yes	4.00 (1.7%)	64.0 (21.5%)	160 (51.3%)	4.00 (1.5%)	65.0 (21.1%)	157 (58.8%)
No	234 (98.3%)	234 (78.5%)	152 (48.7%)	261 (98.5%)	243 (78.9%)	110 (41.2%)
trigmed						
Yes	1.00 (0.4%)	6.00 (2.0%)	18.0 (5.8%)	1.00 (0.4%)	7.00 (2.3%)	12.0 (4.5%)
No	237 (99.6%)	292 (98.0%)	294 (94.2%)	264 (99.6%)	301 (97.7%)	255 (95.5%)
diabetesmed						
Yes	2.00 (0.8%)	22.0 (7.4%)	84.0 (26.9%)	3.00 (1.1%)	14.0 (4.5%)	36.0 (13.5%)
No	236 (99.2%)	276 (92.6%)	228 (73.1%)	262 (98.9%)	294 (95.5%)	231 (86.5%)

[a] Refer to Table 2.5 for the description of each variable

The numerical variables were characterized by the mean, standard deviation (SD), median, minimum (Min) and maximum (Max), by region and overall (Table 4.4).

Table 4.4: Summary of numerical variables by region and overall

Variables [a]	North (N=340)	Center (N=338)	LVT [b] (N=349)	Alentejo (N=343)	Algarve (N=318)	Overall (N=1688)
age						
Mean (SD)	50.0 (18.1)	50.5 (18.1)	50.9 (18.6)	50.3 (18.6)	49.0 (18.2)	50.2 (18.3)
Median [Min, Max]	51.0 [18.0, 79.0]	50.5 [19.0, 79.0]	53.0 [18.0, 79.0]	52.0 [18.0, 79.0]	47.0 [18.0, 79.0]	51.0 [18.0, 79.0]
smokenr						
Mean (SD)	6.76 (11.6)	4.96 (10.8)	8.21 (13.8)	7.32 (12.0)	7.07 (13.0)	6.87 (12.3)
Median [Min, Max]	0 [0, 80.0]	0 [0, 80.0]	0 [0, 80.0]	0 [0, 60.0]	0 [0, 100]	0 [0, 100]
tchol						
Mean (SD)	188 (35.9)	201 (38.1)	191 (38.4)	194 (34.6)	196 (41.2)	194 (37.9)
Median [Min, Max]	188 [87.0, 338]	200 [97.0, 360]	189 [104, 319]	192 [94.0, 329]	191 [94.0, 417]	192 [87.0, 417]
hdl						
Mean (SD)	55.1 (14.2)	57.0 (15.2)	54.9 (16.0)	57.4 (14.8)	56.1 (14.8)	56.1 (15.0)
Median [Min, Max]	53.0 [26.0, 102]	55.0 [22.0, 110]	53.0 [23.0, 112]	56.0 [23.0, 101]	54.0 [23.0, 135]	54.0 [22.0, 135]
ldl						
Mean (SD)	115 (31.3)	127 (35.7)	119 (34.6)	117 (31.3)	120 (36.0)	120 (34.0)
Median [Min, Max]	116 [38.0, 256]	124 [45.0, 274]	115 [45.0, 227]	115 [31.0, 237]	116 [31.0, 296]	117 [31.0, 296]
tg						
Mean (SD)	112 (58.7)	109 (60.8)	111 (69.9)	109 (58.8)	99.5 (51.8)	108 (60.5)
Median [Min, Max]	98.0 [21.0, 461]	96.5 [32.0, 477]	91.0 [27.0, 517]	94.0 [31.0, 450]	86.0 [26.0, 427]	94.0 [21.0, 517]
glucose						
Mean (SD)	95.2 (25.1)	98.9 (22.5)	93.2 (21.2)	99.5 (29.7)	95.5 (23.8)	96.4 (24.7)
Median [Min, Max]	88.5 [61.0, 298]	93.0 [68.0, 203]	89.0 [62.0, 251]	91.0 [70.0, 309]	90.0 [70.0, 287]	91.0 [61.0, 309]
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1.00 (0.3%)	1.00 (0.1%)
sbp						
Mean (SD)	127 (21.6)	134 (23.9)	127 (23.5)	128 (21.5)	124 (20.1)	128 (22.4)
Median [Min, Max]	125 [80.0, 208]	133 [89.5, 203]	124 [80.5, 235]	127 [78.5, 200]	124 [87.0, 223]	127 [78.5, 235]
dbp						
Mean (SD)	81.4 (10.9)	80.9 (11.7)	79.0 (11.6)	79.0 (11.1)	79.2 (11.6)	79.9 (11.4)
Median [Min, Max]	80.5 [56.0, 123]	80.0 [55.0, 119]	78.5 [48.0, 111]	78.5 [47.0, 112]	78.5 [55.5, 133]	79.5 [47.0, 133]
weight						
Mean (SD)	71.8 (13.5)	73.8 (14.4)	74.1 (16.0)	73.5 (15.1)	72.7 (14.5)	73.2 (14.7)
Median [Min, Max]	72.2 [37.5, 120]	72.2 [44.1, 120]	72.0 [40.4, 135]	72.1 [40.4, 129]	71.0 [42.2, 127]	71.9 [37.5, 135]
Missing	1.00 (0.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1.00 (0.1%)
height						
Mean (SD)	1.64 (0.100)	1.63 (0.0874)	1.65 (0.0989)	1.64 (0.0942)	1.65 (0.0993)	1.64 (0.0963)
Median [Min, Max]	1.64 [1.40, 1.95]	1.63 [1.42, 1.85]	1.65 [1.41, 1.92]	1.64 [1.39, 1.96]	1.65 [1.44, 1.93]	1.64 [1.39, 1.96]
bmi						
Mean (SD)	26.4 (4.53)	27.5 (4.85)	26.9 (5.08)	27.1 (5.06)	26.5 (4.74)	26.9 (4.87)
Median [Min, Max]	26.2 [16.2, 40.7]	27.1 [17.1, 45.2]	26.3 [16.8, 55.6]	26.5 [15.9, 42.5]	26.0 [16.1, 46.8]	26.5 [15.9, 55.6]
Missing	1.00 (0.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1.00 (0.1%)
waist						
Mean (SD)	95.0 (11.8)	96.9 (11.9)	91.3 (13.9)	92.0 (14.0)	91.2 (12.1)	93.3 (13.0)
Median [Min, Max]	95.0 [61.0, 131]	97.0 [64.0, 128]	91.5 [60.0, 130]	91.0 [64.0, 138]	91.0 [60.0, 125]	93.0 [60.0, 138]
Missing	0 (0%)	0 (0%)	1.00 (0.3%)	0 (0%)	0 (0%)	1.00 (0.1%)
vpa						
Mean (SD)	0.918 (1.87)	1.03 (1.87)	0.943 (1.74)	1.36 (2.08)	1.57 (2.36)	1.16 (2.00)
Median [Min, Max]	0 [0, 7.00]	0 [0, 7.00]	0 [0, 7.00]	0 [0, 7.00]	0 [0, 7.00]	0 [0, 7.00]
Missing	0 (0%)	0 (0%)	0 (0%)	1.00 (0.3%)	0 (0%)	1.00 (0.1%)
mpa						
Mean (SD)	2.88 (3.02)	4.16 (2.84)	3.29 (3.11)	2.43 (2.60)	3.52 (3.02)	3.25 (2.98)
Median [Min, Max]	2.00 [0, 7.00]	5.00 [0, 7.00]	3.00 [0, 7.00]	1.00 [0, 7.00]	3.00 [0, 7.00]	3.00 [0, 7.00]
Missing	0 (0%)	0 (0%)	0 (0%)	1.00 (0.3%)	0 (0%)	1.00 (0.1%)
mealsday						
Mean (SD)	4.40 (1.05)	4.00 (0.973)	3.93 (1.06)	3.96 (1.21)	4.08 (1.09)	4.07 (1.09)
Median [Min, Max]	4.00 [2.00, 6.00]	4.00 [2.00, 6.00]	4.00 [1.00, 6.00]	4.00 [0, 6.00]	4.00 [0, 6.00]	4.00 [0, 6.00]
wineunits						
Mean (SD)	0.644 (1.05)	1.14 (1.69)	0.688 (1.34)	0.463 (0.806)	0.508 (0.927)	0.691 (1.23)
Median [Min, Max]	0.0300 [0, 10.0]	0.0850 [0, 8.00]	0.0300 [0, 12.0]	0 [0, 4.00]	0.0300 [0, 6.00]	0.0300 [0, 12.0]
beerunits						
Mean (SD)	0.151 (0.441)	0.206 (0.626)	0.210 (0.603)	0.181 (0.690)	0.294 (0.644)	0.207 (0.607)
Median [Min, Max]	0 [0, 3.00]	0 [0, 5.00]	0 [0, 7.00]	0 [0, 10.0]	0.0300 [0, 5.00]	0 [0, 10.0]
whiteunits						
Mean (SD)	0.0443 (0.165)	0.0758 (0.248)	0.0948 (0.352)	0.0472 (0.192)	0.0504 (0.191)	0.0628 (0.241)
Median [Min, Max]	0 [0, 1.00]	0 [0, 2.00]	0 [0, 4.00]	0 [0, 2.00]	0 [0, 2.00]	0 [0, 4.00]
nrfactors						
Mean (SD)	4.20 (2.19)	4.26 (2.20)	4.19 (2.31)	4.20 (2.17)	3.82 (2.09)	4.14 (2.20)
Median [Min, Max]	4.00 [0, 11.0]	4.00 [0, 11.0]	4.00 [0, 11.0]	4.00 [0, 10.0]	4.00 [0, 9.00]	4.00 [0, 11.0]

[a] Refer to Table 2.5 for the description of each variable
[b] Lisbon and Vale do Tejo

As mentioned before, only individuals between the ages of 18 and 79 participated in this study. Regarding the variable **smokenr**, the median is always smaller than the mean, indicating that the distribution of this variable is positively skewed.

For the variables associated with metabolic and physiological parameters as well as physical characteristics there are reference values, indicating whether the observed values are high, normal or low. The ideal value for total cholesterol, **tchol**, is below 200 mg/dL and, in average, that was the case overall and for all regions except for the center region, where the mean for this variable was 201 mg/dL. Levels above 200 mg/dL are considered borderline and above 240 mg/dL are considered very high, and when observing the data, it is possible to see that the maximum values observed within each region and overall exceed 240 mg/dL by far (Figure 4.1). For HDL cholesterol, **hdl**, above 60 mg/dL are considered high, between 40-60 60 mg/dL are considered borderline and below 40 mg/dL are considered low. In average, overall and within each region the values for this variable fall in the borderline category and, the maximum values observed are far larger than 60 and minimum values observed are far smaller than 40 (Figure 4.2). For LDL cholesterol, **ldl**, the normal values range between 100-129, which was verified overall and within each region. Values above 160 mg/dL are very high and individuals with values as such are present in all regions (Figure 4.3). For triglycerides, **tg**, the normal value is below 150 mg/dL and on average, this is the case for all regions and overall. High values, between 200-499 mg/dL, and very high values, above 500 mg/dL, were observed in all regions (Figure 4.4). The normal value for **glucose** is below 110 mg/dL and that was the case for all regions but there were also observed high values for this variable, indicating that some individuals were diabetic (Figure 4.5).

For systolic and diastolic blood pressure, these two variables need to be evaluated simultaneously to determine whether an individual is considered to have hypertension, and that happens when the values for **sbp** are above 140 mmHG and **dbp** above 90 mmHG.

The variables **weight** and **height** were used to determine the **bmi** and when this is above 25 kg/m^2 an individual was pre-obese and if it is above 30 kg/m^2 an individual was considered obese. Looking at the data, on average, the **bmi** indicated pre-obesity since the mean value for this variable as always above 25 kg/m^2 in all regions and overall. There were values far superior to 30 kg/m^2 in all regions (Figure 4.6). The ideal value for the waist perimeter, **waist**, is below 102 cm for men and below 88 cm for women, as can be seen the mean value for this variable was always superior to both values. The variables related to days per week of vigorous and moderated physical activity, **vpa** and **mpa**, ranged between 0 and 7 and the mean value for the variable **mpa** was always superior to the mean value of **vpa**, indicating that on average the individuals practice more moderated forms of physical activity rather than vigorous forms. These variables were proxies for the levels of physical activity. The mean number of meals per day, **mealsday**, was around 4 in all regions and overall, registering the minimum value of 0 and maximum of 6. This variable was a proxy of the quality of one's diet. The variables related to alcohol consumption - **wineunits**, **beerunits** and **whiteunits** – had different mean values and wine seems to be the most consumed alcohol beverage, followed by beer and white beverages being the least consumed. It was considered that 1 unit of each beverage corresponded to 10 grams of alcohol. The consumption of more than 10 gr of alcohol for women and more than 20 gr for men is considered an unhealthy alcohol consumption.

The mean value for the number of risk factors each individual presented, **nrfactors**, was around 4 in each region and overall. There were individuals with 0 risk factors and no individual presented all the risk factors considered.

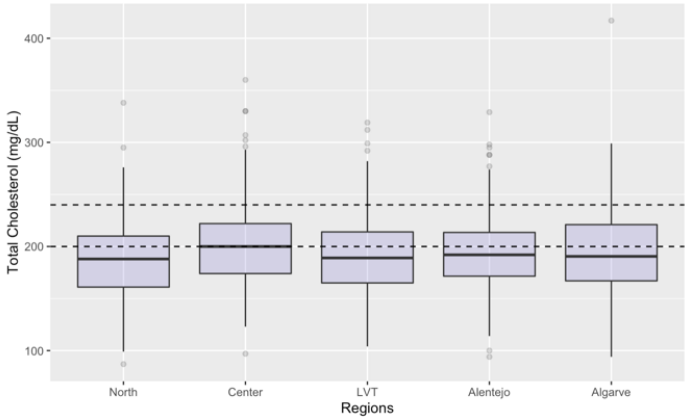

Figure 4.1: Distribution of total cholesterol values by region. Horizontal lines representing borderline and very high values according to clinical criteria

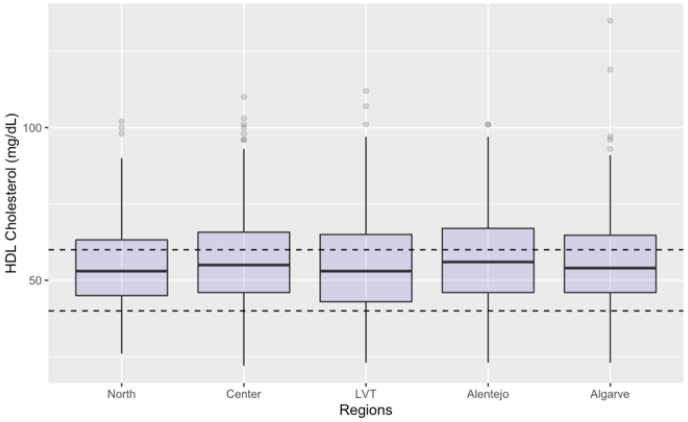

Figure 4.2: Distribution of HDL-cholesterol values by region. Horizontal lines representing low and borderline values according to clinical criteria.

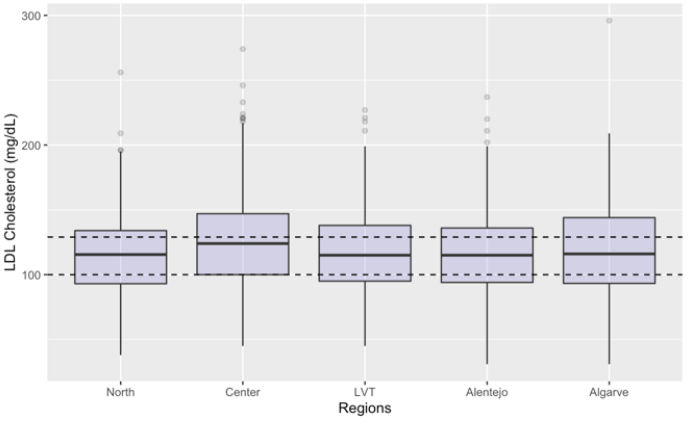

Figure 4.3: Distribution of LDL-cholesterol values by region. Horizontal lines representing normal range of values according to clinical criteria

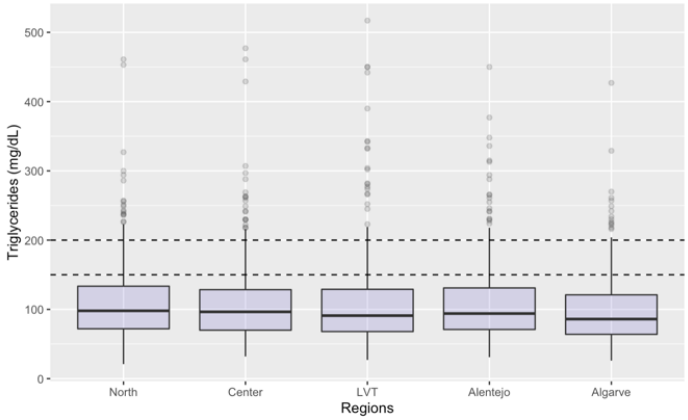

Figure 4.4: Distribution of triglyceride levels by region. Horizontal lines representing borderline and high values according to clinical criteria.

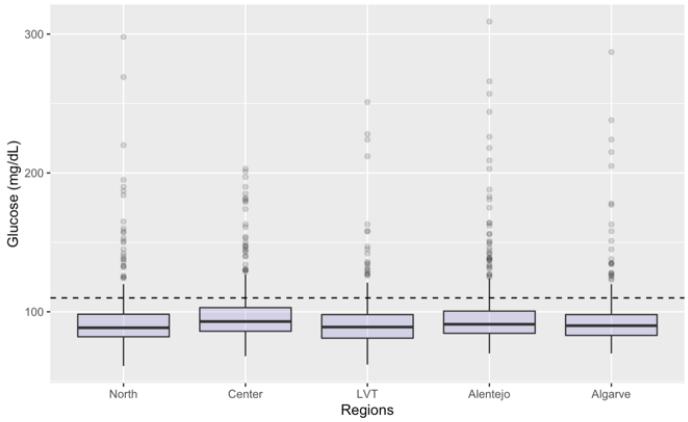

Figure 4.5: Distribution of glucose levels by region. Horizontal line representing high values, according to clinical criteria.

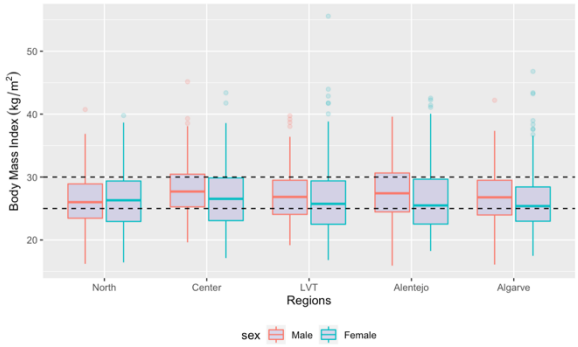

Figure 4.6: Distribution of body mass index values by region and sex. Horizontal lines representing BMI values that indicate pre-obesity and obesity, according to clinical criteria.

Data from INE regarding healthcare access and data from the Portuguese adapted EDI was analyzed and the results are summarized in Table 4.5. It is possible to observe differences between the regions, having really deprived regions such as Algarve where all the municipalities sampled belonged to the 5th quintile of the EDI, the most deprived one, contrasting with the Center region, whose municipalities belonged to the 1st and 2nd quintile, the least deprived ones. It is also interesting to see that in the North region, there are two municipalities with very different levels of deprivation, Porto

belonging to the most deprived quintile and Maia belonging to the 3rd quintile, which is right in the middle of the EDI.

The municipality with the highest number of healthcare centers was Lisbon, having 17, followed by Porto having 8 healthcare centers. This makes sense considering these two municipalities are the most populated ones. The municipality with the highest number of doctors was also Porto, with 12, followed by Montemor-o-Velho, with 10 doctors per 10 000 habitants. The municipality with the highest number of nurses per 10 000 habitants was Montemor-o-Novo, with 13 nurses, followed by Ponte de Sor with 12 nurses per 10 000 habitants. The municipality with the least number of doctors was Ponte de Sor, with 4, and with the least number of nurses was Odivelas and Pombal, with 5 each.

Table 4.5: Summary of socioeconomic and healthcare access related variables

Region	Municipality	EDI Score	EDI Quintile	Number of Healthcare Centers	Number of Doctors[a]	Number of Nurses[b]
North	Porto	2,7496	5	8	12	11
	Maia	-0,8975	3	3	7	7
Center	Pombal	-3,8559	1	1	6	5
	Montemor-o-Velho	-1,9585	2	1	10	6
	Soure	-2,4533	1	1	9	10
LVT[c]	Odivelas	3,8277	5	2	5	5
	Lisbon	2,0280	4	17	8	6
Alentejo	Montemor-o-Novo	3,6733	5	1	8	13
	Évora	0,5889	4	1	9	7
	Ponte de Sor	1,6044	4	2	4	12
Algarve	Faro	3,6556	5	1	9	9
	Olhão	3,9141	5	1	6	8
	Silves	5,3567	5	1	6	8
	Portimão	4,3847	5	1	6	7

[a] Number of doctors per 10 000 habitants, calculated as the quotient between the number of doctors in the municipality and the resident population times 10 000.
[b] Number of nurses per 10 000 habitants, calculated as the quotient between the number of nurses in the municipality and the resident population times 10 000.
[c] Lisbon and Vale do Tejo

4.2 Modelling the occurrence of past cerebro-cardiovascular disease – original data

4.2.1 Pre-selection of variables

Due to having a high number of variables available it was decided to do a pre-selection of variables before beginning the stepwise selection process. There was the possibility that some of the variables available could be correlated. One of the variables collected was body mass index and that is a measure of body fat that takes in consideration a person's weight and height (Equation 4.1). The waist perimeter, or waist circumference, is also a measure of body fat, consequently affected by a person's weight. So, it made sense to determine whether there was a correlation between the variables **weight**, **height**, **bmi** and **waist**. It was found a high degree of correlation between the variables **bmi** and **waist**

($\rho = 0.86$), between **bmi** and **weight** ($\rho = 0.81$) and between **weight** and **waist** ($\rho = 0.80$), these results are illustrated in Figure 4.7.

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2} \tag{4.1}$$

Among the metabolic parameters collected, there could be some correlations as well. Total cholesterol is a measure of the total amount of cholesterol in the blood, including low-density lipoprotein (LDL), commonly known as "bad cholesterol", high-density lipoprotein (HDL), commonly known as "good cholesterol" and it also takes into consideration the levels of triglycerides, a very common type of fat present in the body (Equation 4.2). The data demonstrated a high degree of correlation between the variables **tchol** and **ldl** ($\rho = 0.93$), as can be seen in Figure 4.8.

$$\text{Total Cholesterol} = \text{HDL level} + \text{LDL level} + 0.20 \times \text{Triglyceride level} \tag{4.2}$$

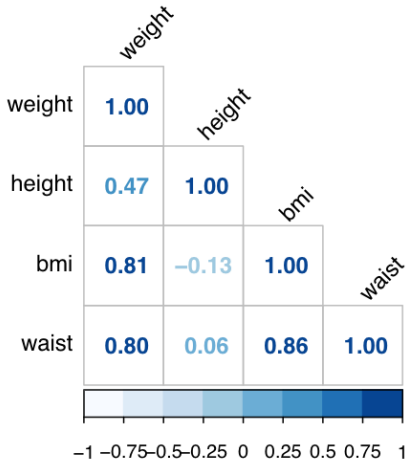

Figure 4.7: Correlation coefficient between physical characteristics

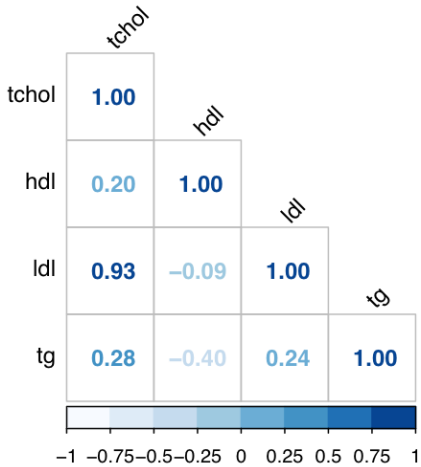

Figure 4.8: Correlation coefficient between metabolic parameters

Determining the correlation coefficient is not enough to decide whether to put a variable up for selection for inclusion in the model, so the usefulness of these variables was also evaluated creating univariate logistic regression models, for each of the previously mentioned covariables. All variables, except for **tg,** were found to be significant at the usual level of significance of 0.05, the results found are displayed in Table 4.6. The notation used in this table will be the same going forward in this work and "*Y*~*X*" means that the variable *Y* is the response variable and is regressed on the covariable *X*.

Table 4.6: Summarized results from univariate logistic regression models

Univariate Models	Estimate	p-value	OR	95% CI
cvevent ~ weight	0.020	0.004	1.023	(1.006, 1.035)
cvevent ~ height	-2.609	0.029	0.074	(0.007, 0.746)
cvevent ~ bmi	0.095	< 0.001	1.100	(1.056, 1.144)
cvevent ~ waist	0.052	< 0.001	1.053	(1.035, 1.070)
cvevent ~ tchol	-0.016	< 0.001	0.984	(0.978, 0.991)
cvevent ~ hdl	-0.020	0.016	0.981	(0.965, 0.996)
cvevent ~ ldl	-0.018	< 0.001	0.982	(0.974, 0.989)
cvevent ~ tg	0.003	0.11	1.003	(0.999, 1.006)

OR: odds ratio; CI: confidence interval

In addition to this, the variance inflation factor was also determined. To do this, the response variable, **cvevent,** was regressed on the variables **weight**, **height**, **bmi** and **waist** and then the VIF of the variables included in the model was determined. It was found that the variable **waist** was the least correlated with the rest and **weight** was the most correlated, having the highest VIF (Table 4.7). The response variable was once again regressed on the same variables expect for the one with the highest VIF, **weight**, and the new VIFs were determined, being now all below 5, meaning that there is no correlation between the variables in the model (Table 4.8).

Table 4.7: Determined Variance Inflation Factors (VIF) for the model with variables weight, height, bmi and waist

Variable	weight	height	bmi	waist
VIF	105.54	38.10	76.90	3.97

Table 4.8: Determined Variance Inflation Factors (VIF) for the model with variables height, bmi and waist

Variable	height	bmi	waist
VIF	1.12	3.93	3.88

This same process was repeated for the variables **tchol, hdl, ldl** and **tg,** and **tchol** was the variable with the highest VIF. Like before, it was then removed from the model and the VIFs were recalculated for the model without **tchol**, and this time all were below 5, indicating no correlation between the variables (Tables 4.9 and 4.10).

Table 4.9: Determined Variance Inflation Factors (VIF) for the model with variables tchol, hdl, ldl and tg

Variable	tchol	hdl	ldl	tg
VIF	48.63	9.28	38.93	5.67

Table 4.10: Determined Variance Inflation Factors (VIF) for the model with variables hdl, ldl and tg

Variable	**hdl**	**ldl**	**tg**
VIF	1.22	1.03	1.25

From this previous analysis it was concluded that there was no need to continue further with the variables **weight** and **tchol**, since the data indicated that these were correlated with other variables available, and it would be redundant including them in the same model.

The interactions between the biological and physiological parameters and medication that affects those parameters were also considered as possible predictors to be included in the model (hdl*cholmed, ldl*cholmed, tg*trigmed, glucose*diabetesmed, sbp*hypertensorsmed and dbp*hypertensorsmed). Even though these interactions were not found to be statistically significant (p-values were above 0.05) they are interesting to include since medication influences the values the biological and physiological parameters take.

4.2.2 Stepwise Selection

The process of stepwise selection was used to obtain a model including biological characteristics of the individuals as well as lifestyle habits and variables related to inequalities. The first step was to obtain the base model, with the response variable **cvevent** regressed on the variables **age** and **sex** (Equation 4.3). These two covariables are considered confounding factors because both are risk factors for the outcome, and will always be present in the model, regardless of their p-value.

$$\text{logit}(\text{p}) = -7.016 - 0.431\ \text{Female} + 0.072\ \text{Age} \tag{4.3}$$

When interpreting this model, it is possible to conclude that the odds of having had a cerebro-cardiovascular event for a male individual at age zero corresponds to the exponential of $\widehat{\beta_0}$, that is 0.0009. The fitted model shows that, holding age at a fixed value, the odds of having had a cerebro-cardiovascular event for females (sex = 1) over the odds of having had a cerebro-cardiovascular event for males (sex = 0) corresponds to the exponential of $\widehat{\beta_1}$, that is 0.65. The estimated coefficient for age shows that holding sex at a fixed value, we will see an increase of 7% in the odds of having had a cerebro-cardiovascular event, for a one-unit increase in age since the exponential of $\widehat{\beta_2}$ is 1.0743. Regarding the p-values, the estimates of the intercept and age are significant for all the usual significance levels. However, the p-value for the estimate of the variable sex is above the significance level of 0.05. All the estimates from the model are summarized in Table 4.11.

Table 4.11: Summarized results from the model with the covariables sex and age

	Estimate	**p-value**	**OR**	**95% CI**
sex				
female	-0.431	0.068	0.650	(0.406, 1.027)
age	0.072	<0.001	1.074	(1.055, 1.097)

OR: odds ratio; CI: confidence interval

After this, the response variable was regressed on the variables age, sex and all the other possible predictors individually to find the one with the smallest p-value, inferior to the inclusion p-value (0.20), to be added to the model. The variable with the smallest p-value was **cholmed** (0,0000005), so this was

the first variable added to the base model. The results from the first step of the stepwise selection are presented in Table 4.12.

Table 4.12: Summary of the first step in the stepwise selection method

Model Formula	$p - value_{\hat{\beta}_3}$
cvevent ~ sex + age + hdl	0.063
cvevent ~ sex + age + hdl * cholmed	0.704
cvevent ~ sex + age + ldl	<0.001
cvevent ~ sex + age + ldl * cholmed	0.527
cvevent ~ sex + age + tg	0.903
cvevent ~ sex + age + tg * trigmed	0.682
cvevent ~ sex + age + glucose	0.119
cvevent ~ sex + age + glucose * diabetesmed	0.717
cvevent ~ sex + age + sbp	0.516
cvevent ~ sex + age + sbp * hypertensorsmed	0.538
cvevent ~ sex + age + dbp	0.335
cvevent ~ sex + age + dbp * hypertensorsmed	0.636
cvevent ~ sex + age + height	0.516
cvevent ~ sex + age + bmi	0.027
cvevent ~ sex + age + waist	0.026
cvevent ~ sex + age + cholmed	<0.001
cvevent ~ sex + age + trigmed	0.010
cvevent ~ sex + age + diabetesmed	0.005
cvevent ~ sex + age + hypertensorsmed	0.0002
cvevent ~ sex + age + smokenr	0.0002
cvevent ~ sex + age + mealsday	0.807
cvevent ~ sex + age + dieteq	0.508
cvevent ~ sex + age + wineunits	0.362
cvevent ~ sex + age + beerunits	0.452
cvevent ~ sex + age + whiteunits	0.791
cvevent ~ sex + age + alcoholpf	0.812
cvevent ~ sex + age + palevelModerate	0.193
cvevent ~ sex + age + palevelLow	0.408
cvevent ~ sex + age + vpa	0.085
cvevent ~ sex + age + mpa	0.849
cvevent ~ sex + age + historyrf	<0.001

Every time a new variable was added to the model, all the p-values of the other variables already present, apart from sex and age, were evaluated to make sure none were superior to the exclusion p-value, if that were the case, said variable would be excluded. This process was repeated until there were no variables with p-value inferior to the inclusion p-value. Once that happened, the variables in the model that had a p-value superior to the significance level of 0.05 were removed from the model, in descending order of p-value, because even though said p-value was inferior to the inclusion p-value, the

variable cannot be considered statistically significant. Table 4.13 has a summary of the stepwise selection process performed to reach the final model.

Table 4.13: Summary of all the steps in the stepwise selection process

Step	Model Formula[a]	AIC
1st	Base model + cholmed	558.79
2nd	1st model + historyrf	542.24
3rd	2nd model + smokenr	529.66
4th	3rd model + ldl	519.18
5th	4th model + hypertensorsmed	516.05
6th	5th model + trigmed	516.27
7th	6th model + bmi	516.48
8th	7th model - bmi	516.27
9th	8th model - trigmed	516.05

AIC: Akaike Information Criterion
[a] refer to Table 1 in appendices

As can be seen in the previous table, the AIC lowers as the new variables are entered in the model until the 6th step where it suffers a slight increase as well as in the 7th step. The final model, obtained in step 8 (same as the one in step 5), has the lowest AIC, meaning that is the better-fit model. The model obtained at the end of the stepwise selection process included the covariables **sex**, **age**, **cholmed**, **historyrf**, **smokenr**, **ldl** and **hypertensorsmed**. The next step was adding the variables that can reflect inequalities such as academic qualifications, **education**, and deprivation level, **ediscore**. The random effect of the **region** was also included in the model. The final model can be written through the following simplified expression:

$$\begin{aligned}\text{logit}(p) = -5.165 - 0.128\ \text{sexF} + 0.039\ \text{Age} + 1.102\ \text{cholmedY} + 1.462\ \text{historyrf1} + \\ 0.029\ \text{smokenr} - 0.013\ \text{ldl} + 0.698\ \text{hypertensorsmedY} - 0.306\ \text{education2} - \\ 0.518\ \text{education3} + 0.052\ \text{ediscore} + a\end{aligned} \tag{4.4}$$

where *a* represents the random effect of the region. This expression can be written in this simplified way, without indexes, considering that doing so does not affect its understanding. The results from this final model are summarized in Table 4.14.

Table 4.14: Summarized results from the final model

	Estimate	p-value	OR	95% CI
age	0.039	0.002	1.040	(1.015, 1.065)
smokenr	0.029	<0.001	1.029	(1.014, 1.044)
ldl	−0.013	0.001	0.987	(0.979, 0.995)
ediscore	0.052	0.320	1.053	(0.936, 1.171)
sex				
female	−0.128	0.652	0.880	(1.015, 1.065)
cholmed				
yes	1.102	<0.001	3.010	(1.660, 5.628)
hypertensorsmed				
yes	0.698	0.042	2.011	(1.049, 4.064)
historyrf				

yes	1.462	<0.001	4.316	(2.222, 8.130)
education				
high school	− 0.306	0.413	0.737	(0.339, 1.487)
university education	−0.518	0.322	0.596	(0.190, 1.535)

OR: odds ratio; CI: confidence interval

The interpretation of the model is as follows:

- **age**: when this variable suffers a one unit increase, the odds of the outcome are multiplied by the exponential of its estimate, that is, exp (0.039) = 1.040. When holding all the other variables at a fixed value, we will see an increase of 4% in the odds of having had a cerebro-cardiovascular event.
- **smokenr**: when this variable suffers a one unit increase, the odds of the outcome are multiplied by the exponential of its estimate, that is, exp (0.029) = 1.029. When holding all the other variables at a fixed value, we will see an increase of 3% in the odds of having had a cerebro-cardiovascular event.
- **ldl**: when this variable suffers a one unit increase, the odds of the outcome are multiplied by the exponential of its estimate, that is, exp (-0.013) = 0.987 When holding all the other variables at a fixed value, we will see a decrease in the odds of having had a cerebro-cardiovascular event. Usually, when an individual has had a cerebro-cardiovascular event they will be medicated to control cholesterol levels in the blood. So, people that have had a past cerebro-cardiovascular event will most likely present lower levels of LDL-c, due to it being controlled by medication, hence why the negative estimate for this variable's coefficient.
- **ediscore**: when this variable suffers a one unit increase, the odds of the outcome are multiplied by the exponential of its estimate, that is, exp (0.052) = 1.053. When holding all the other variables at a fixed value, we will see an increase of 5% in the odds of having had a cerebro-cardiovascular event.
- **sex**: the odds of having had a cerebro-cardiovascular event for females (sex = 1) over the odds of having had a cerebro-cardiovascular event for males (sex = 0) corresponds to the exponential of the estimate for sex, that is, exp (-0.128) = 0.880, when holding all the other variables at a fixed value. The fact that the estimate is negative means that being a female is associated with a reduction in the odds of having had the event.
- **cholmed**: the odds of having had a cerebro-cardiovascular event for individuals medicated for cholesterol (cholmed = yes) is 3 times that of the individuals not medicated (cholmed = no), corresponding to the exponential of the estimate, that is, exp (1.102) = 3.010, when controlling for all other variables. This makes sense considering that individuals that have had a CCV event will most likely be medicated to control their levels of cholesterol.
- **hypertensorsmed**: the odds of having had a CCV event for individuals medicated for hypertension (hypertensorsmed = yes) is 2 times that of the individuals not medicated, (hypertensorsmed = no) when controlling for all other variables. This makes sense because individuals that have had a CCV might be medicated for hypertension.
- **historyrf**: an individual that has a history of cerebro-cardiovascular disease (historyrf = yes) has 4.3 times the odds of having had a CCV event than an individual that does not have such history (historyrf = no), when controlling for all other variables.
- **education**: this variable is categorical, and the reference category was "none/pre–high school qualifications". The odds of having had a cerebro-cardiovascular event for an individual that completed High School (education = 2) over the odds of having had a cerebro-cardiovascular event for individuals with none/pre–high school qualifications corresponds to 0.737. The odds of having had a cerebro-cardiovascular event for an individual that completed University Education

(education = 3) over the odds of having had a cerebro-cardiovascular event for individuals none/pre–high school qualifications corresponds to 0.596. The fact that the estimates for the categories of this variable are negative means that they are associated with a reduction in the odds of having had the event, i.e., higher education levels reduce the odds of the outcome.

When looking at the p-values, it is possible to conclude that the variables related to inequalities are not statistically significant since their p-values exceed the significance level of 0.05. This indicates that with all the previous information already present in the model, knowing an individual's academic qualifications or level of deprivation does not seem to improve the ability to predict whether they had a previous cerebro-cardiovascular event.

Variables with an odds ratio approximately equal to one do not really have an effect on the odds of the outcome, and that is the case for the variables **age**, **smokenr**, **ldl** and **ediscore**. Variables with an odds ratio above one are associated with an increase in the odds of the outcome, happening with variables **cholmed**, **hypertensorsmed** and **historyrf**. Variables with an odds ratio inferior to one are associated with lower odds of the outcome, in this case that happens for the variable sex and with the variable education (Figure 4.9).

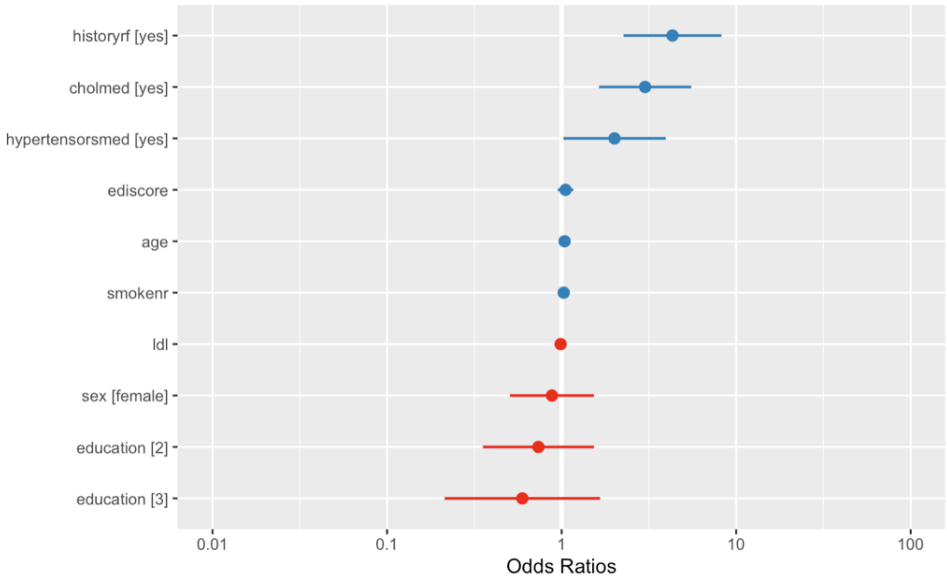

Figure 4.9: Odds ratios of the final model's fixed effects with 95% confidence intervals. Education2 corresponds to "High School" and education3 corresponds to "University Education".

The random effect of the region was also analyzed, and it appears to not have a big impact since the conditional modes take values very close to zero and all the confidence intervals overlap each other which indicates the effect is not statistically significant (Figure 4.10).

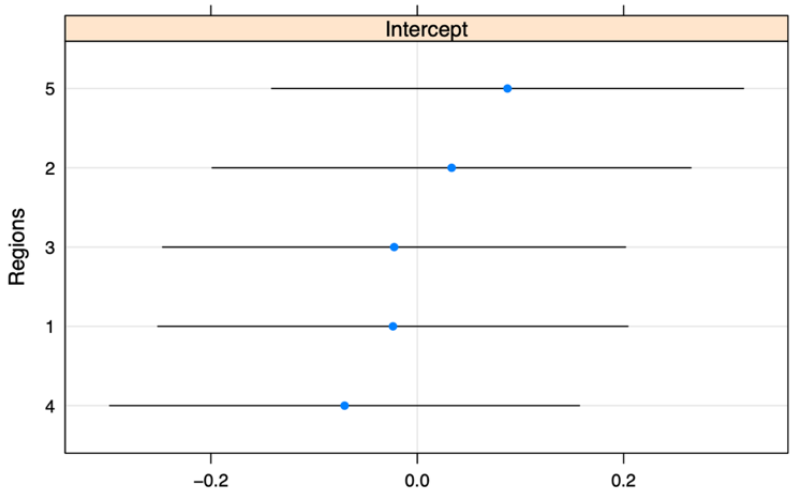

Figure 4.10: Region random effects given by conditional modes with 95% confidence intervals based on the conditional variance. Regions 1-5 correspond to North, Center, Lisbon and Vale do Tejo, Alentejo and Algarve, respectively.

4.2.3 ROC Curve

The ROC curve obtained had an AUC of 0.870 which suggests that 87% of the time the model will correctly distinguish a case from a non-case (Figure 4.11). As mentioned previously in chapter 3, an AUC between 0.8 to 0.9 is considered excellent. Specificity and sensitivity are specific to each selected threshold. The optimal threshold is the one that maximizes the sum of sensitivity and specificity. In this case, the optimal threshold was 0.070.

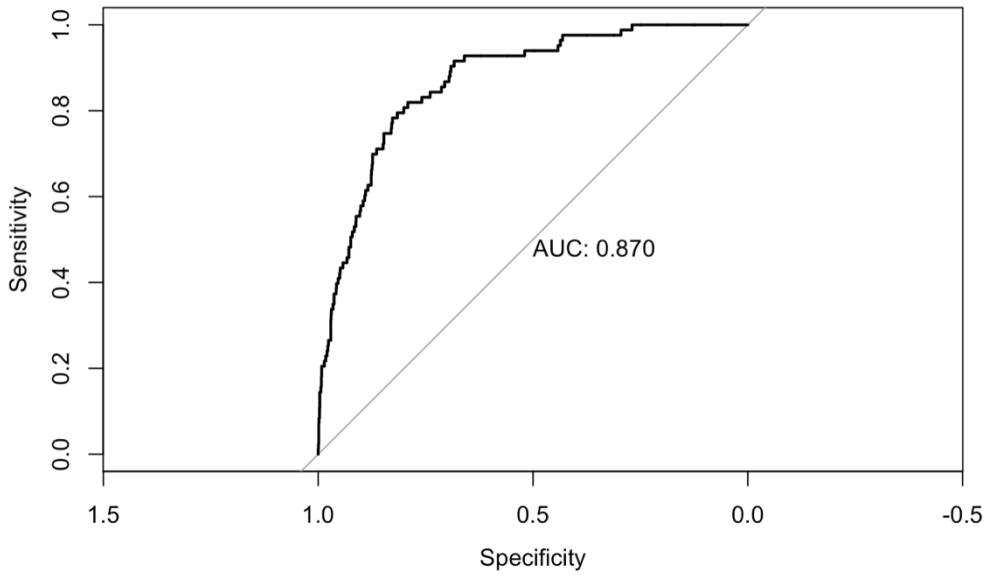

Figure 4.11: ROC curve with respective AUC

4.2.4 Cross validation

To conduct the cross-validation procedures the model used was the model without the random effect of the region (Equation 4.5).

$$\begin{aligned}\text{logit(p)} = \widehat{\beta_0} + \widehat{\beta_1}\,\text{sexF} + \widehat{\beta_2}\,\text{age} + \widehat{\beta_3}\,\text{cholmedY} + \widehat{\beta_4}\text{historyrf1} + \widehat{\beta_5}\text{smokenr} + \widehat{\beta_6}\text{ldl} \\ + \widehat{\beta_7}\text{hypertensorsmedY} + \widehat{\beta_8}\text{education2} + \widehat{\beta_9}\,\text{education3} \\ + \widehat{\beta_{10}}\,\text{ediscore}\end{aligned} \quad (4.5)$$

When using the 10-fold approach, the estimated mean value for accuracy was 0.952, meaning that about 95% of all instances will be correctly classified when using this model. The mean value for Cohen's Kappa will not be considered because, as it was mentioned before in chapter 3, Kappa is not the best metric to evaluate model performance when there is an imbalanced distribution of classes.

The accuracy in each fold was about 95%. The 10-fold cross validation estimate for the test error is approximately 0.041.

Table 4.15: 10-fold cross validation results

Accuracy	Accuracy SD
0.952	0.005

SD: standard deviation

When using the leave-one-out approach, the estimated mean value for accuracy was 0.952, meaning that about 95% of all instances will be correctly classified when using this model and the estimate for the test error is approximately 0.041. Both approaches have very similar results.

4.3 Modelling the occurrence of past cerebro-cardiovascular disease – balanced data

In the original data set, there were 1688 individuals and only 84 had a past cerebro-cardiovascular event. In order to obtain a more balanced data set, the SMOTE technique was applied, like explained in the previous chapter. Said technique resulted in a data set with 840 individuals, 420 that had a previous cerebro-cardiovascular event and 420 that did not.

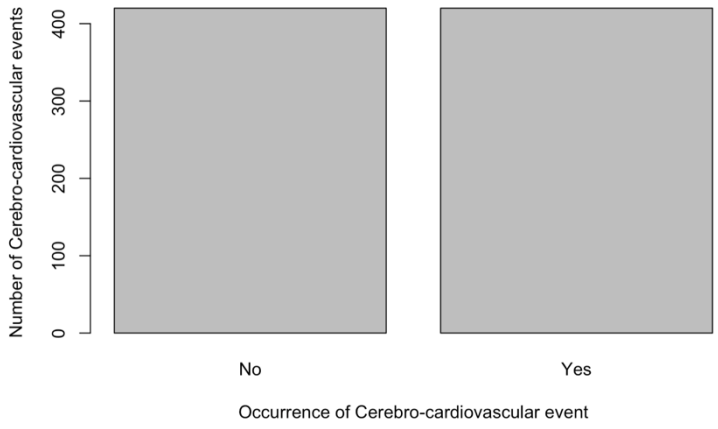

Figure 4.12: Proportion of individuals with and without a past cerebro-cardiovascular event

4.3.1 Pre-selection of variables

The strategy used for the original dataset was the same used for the dataset obtained through the SMOTE technique. The same variables were tested for correlation and multicollinearity and very similar results were obtained, proceeding to the stepwise selection process with the exact same variables as before, that is, excluding **tchol** and **weight** from the analysis.

4.3.2 Stepwise Selection

Once again, the base model had the response variable **cvevent** regressed on the covariables that constitute confounding factors, **sex** and **age** (Equation 4.6). The results of the estimation of the model are summarized in Table 4.16.

$$\text{logit}(p) = -6.905 - 1.284\ \text{Female} + 0.115\ \text{Age} \tag{4.6}$$

Table 4.16: Summarized results from the model with the covariables sex and age

	Estimate	p-value	OR	95% CI
sex				
female	-1.284	<0.001	0.277	(0.191, 0.400)
age	0.115	<0.001	1.122	(1.102, 1.144)

OR: odds ratio; CI: confidence interval

After the base model was established, it was time to start the stepwise selection process for this balanced data set, using the same approach as before. This stepwise selection process had 24 steps, the first variable to enter the model was **cholmed** with a very small p-value (<0.001). On the 21st step there were no more variables with p-value inferior to the inclusion p-value and there were no variables with p-value superior to the exclusion p-value. However, some of the variables in this model had p-value larger than 0.05 and were removed (Table 4.17).

Table 4.17: Summary of Stepwise Selection Process

Step	Model	AIC
1st	Base model + cholmed	694.38
2nd	1st model + historyrf	650.40
3rd	2nd model + waist	622.86
4th	3rd model + ldl	601.53
5th	4th model + trigmed	579.83
6th	5th model + smokenr	570.29
7th	6th model + tg*trigmed	556.78
8th	7th model + beerunits	554.57
9th	8th model + glucose*diabetesmed	555.79
10th	9th model + dieteq	553.47
11th	10th model + mealsday	551.08
12th	11th model + diabetesmed	551.08
13th	12th model + vpa	548.97

14th	13th model + hypertensorsmed	547.95
15th	14th model + sbp	546.37
16th	15th model + glucose	546.37
17th	16th model + mpa	545.65
18th	17th model + wineunits	545.51
19th	18th model + alcoholpf	540.44
20th	19th model + dbp	540.57
21st	20th model - dbp	540.44
22nd	21st model - mpa	541.00
23rd	22nd model - sbp	541.85
24th	23rd model - glucose	541.85

AIC: Akaike information criterion
[a] refer to Table 2 in appendices

After all the non-significant variables were removed, the variables that can reflect inequalities such as academic qualifications, **education**, and deprivation level, **ediscore**, were added to the model. The random effect of the **region** was also included in the model (Equation 4.7).

$$\begin{aligned}\text{logit(p)} = \; & -7.497 - 0.736\,\text{sexF} + 0.036\,\text{age} + 1.541\,\text{cholmedY} \\ & + 2.057\,\text{historyrfY} + 0.059\,\text{waist} - 0.016\,\text{ldl} + 0.039\,\text{smokenr} \\ & - 0.020\,\text{tg} \times \text{trigmedY} - 0.878\,\text{beerunits} \\ & - 0.015\,\text{glucose} \times \text{diabetesmedY} + 0.614\,\text{dieteqY} \\ & - \; 0.195\,\text{mealsday} - 0.213\,\text{vpa} + 0.527\,\text{hypertensorsmedY} \\ & - 0.431\,\text{wineunits} - 0.950\,\text{alcoholpfY} - 0.269\,\text{education2} \\ & - 0.742\,\text{education3} + 0.083\,\text{ediscore} + a\end{aligned} \tag{4.7}$$

where *a* represents the random effect of the region. Education2 represents "High School" and education3 represents "University Education". The results from this final model are summarized in Table 4.18.

Table 4.18: Summarized results from the final model

Variables	Estimate	p-value	OR	95% CI
sex female	-0.736	0.013	0.479	(0.268, 0.857)
age	0.036	0.008	1.037	(1.010, 1.066)
cholmed yes	1.541	<0.001	4.667	(2.636, 8.263)
historyrf yes	2.057	<0.001	7.820	(3.610, 16.940)
waist	0.059	<0.001	1.061	(1.036, 1.086)
ldl	-0.016	<0.001	0.984	(0.975, 0.993)
smokenr	0.039	<0.001	1.040	(1.019, 1.062)
tg*trigmed	-0.020	0.001	0.981	(0.969, 0.993)
beerunits	-0.878	0.001	0.416	(0.243, 0.713)
glucose*diabetesmed	-0.015	0.128	0.985	(0.966, 1.004)
dieteq yes	0.614	0.019	1.848	(1.107, 3.085)
mealsday	-0.195	0.080	0.823	(0.662, 1.024)
vpa	-0.213	0.021	0.808	(0.674, 0.969)

hypertensorsmed				
yes	0.527	0.062	1.695	(0.974, 2.949)
wineunits	-0.431	0.002	0.650	(0.493, 0.858)
alcoholpf				
yes	-0.950	0.007	0.387	(0.194, 0.770)
education				
high school	-0.269	0.448	0.764	(0.382, 1.531)
university education	-0.742	0.202	0.476	(0.153, 1.487)
ediscore	0.083	0.235	1.087	(0.947, 1.247)

OR: odds ratio; CI: confidence interval

The interpretation of the estimates of this model is done the same way as for the previously obtained model for the original data set, since both are generalized linear mixed models. Once again, the variables **ediscore** and **education** are not statistically significant, with p-values above 0.05.

When comparing the two models, the one for the original data set and this one, for the balanced data set, the later one includes all the same variables as the first one plus the following variables: **waist**, **tg*trigmed** (variable that represents the interaction between the levels of triglycerides and triglycerides medication), **beerunits**, **glucose*diabetesmed** (variable that represents the interaction between the levels of glucose and diabetes medication), **dieteq**, **mealsday**, **vpa**, **wineunits** and **alcoholpf**.

For the common variables, the ones that stood out for having a big increase in the odds ratio were **cholmed,** with an increase in the OR from 3.010 to 4.667 and **historyrf**, with an increase in the OR from 4.316 to 7.820. This means that in this data set, an individual that has a history of cerebro-cardiovascular disease has 7.8 times the odds of having had a CCV event than an individual that does not have such history, while holding all other variables constant, and the odds of having had a CCV event for individuals medicated for cholesterol is 4.7 times that of the individuals not medicated, when controlling for all other variables.

The variables **cholmed**, **historyrf**, **hypertensorsmed** and **dieteq** have an OR larger than one, meaning that they are associated with an increase in the odds of the outcome (Figure 4.13).

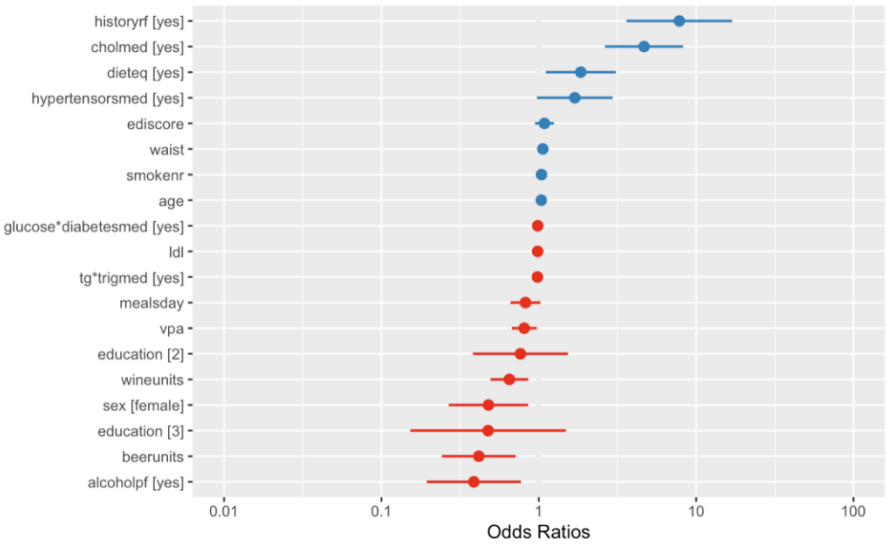

Figure 4.13: Odds ratios of the final model's fixed effects with 95% confidence intervals. Education2 corresponds to "High School" and education3 corresponds to "University Education".

The random effect of the region was once again analyzed, and with this model in this data set the region has a greater effect than it did before. However, all the confidence intervals overlap each other which indicates the effect is not statistically significant (Figure 4.14).

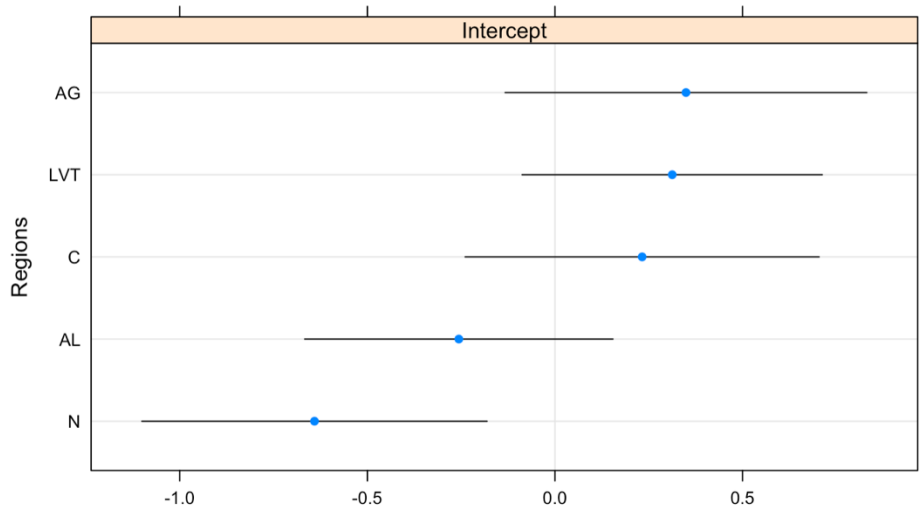

Figure 4.14: Region random effects given by conditional modes with 95% confidence intervals based on the conditional variance. AG - Algarve, LVT - Lisbon and Vale do Tejo, C -Center, AL – Alentejo, N – North.

4.3.3 ROC Curve

The ROC curve obtained had an AUC of 0.950 which suggests that 95% of the time the model will correctly distinguish a case from a non-case (Figure 4.15). As mentioned previously in chapter 3, an AUC of more than 0.9 is considered outstanding. Specificity and sensitivity are specific to each selected threshold. The optimal threshold is the one that maximizes the sum of sensitivity and specificity. In this case, the optimal threshold was 0.545.

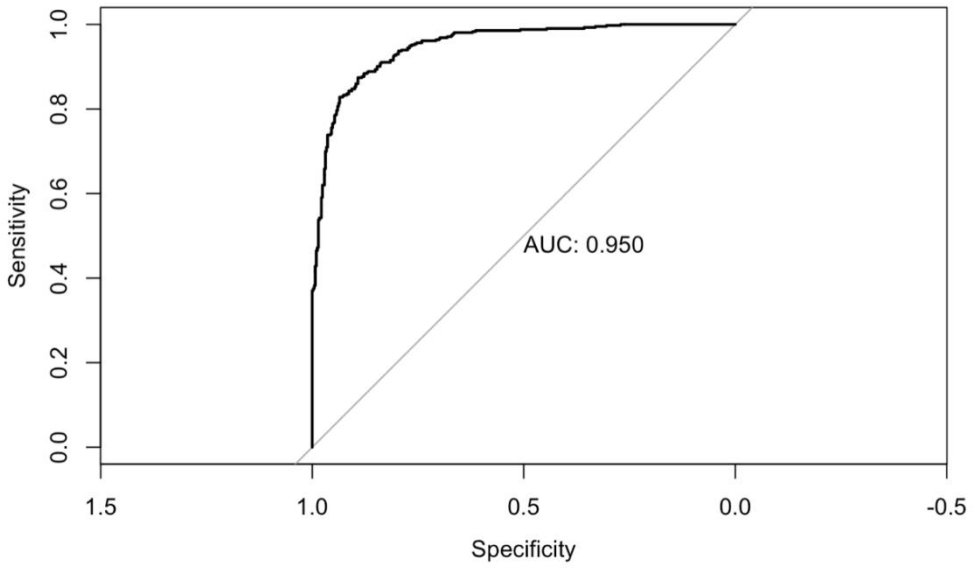

Figure 4.15: ROC curve with respective AUC.

4.3.4 Cross validation

To conduct the cross-validation procedures, the model used was the model without the random effect of the region.

When using the 10-fold approach, the estimated mean value for accuracy was 0.858, meaning that about 86% of all instances will be correctly classified when using this model. The mean value for Cohen's Kappa was also obtained, with the value 0.716 (Table 4.19). A Kappa between 0.60-0.79 represents moderate level of agreement between the classification produced by the model and the real classification. The accuracy in each fold varied between 80-92%. The 10-fold cross validation estimate for the test error was approximately 0.103.

Table 4.19: 10-fold cross validation results

Accuracy	Kappa	Accuracy SD	Kappa SD
0.858	0.716	0.038	0.076

SD: standard deviation

When using the leave-one-out approach, the estimated mean value for accuracy was 0.851, meaning that about 85% of all instances will be correctly classified when using this model. The mean value for Cohen's Kappa was also obtained, with the value 0.701, suggesting a moderate level of agreement between the classification produced by the model and the real classification. The leave-one-out cross validation estimate for the test error was approximately 0.102. Both approaches had very similar results, with accuracy around 85% and Kappa approximately 0.70.

4.4 Modelling the number of risk factors

The number of risk factors was the other response variable of interest in this work. This variable, **nrfactors**, was created by doing the sum of the risk factors each individual presented. The method used to model the response variable was Poisson regression. The risk factors for cerebro-cardiovascular disease considered were:

- High BMI (**bmirf**)
- High Waist Perimeter (**waistrf**)
- Hypertension (**hypertensionrf**)
- LDL-cholesterol > 160 mg/dL (**ldl160**)
- HDL-cholesterol < 40 mg/dL (**hdlrf**)
- Total cholesterol > 240 mg/dL (**hct240**)
- Triglycerides > 200 mg/dL (**htg200**)
- History of CVD (**historyrf**)
- Diabetes (**diabetesrf**)
- Smoking habits (**smokingrf**)
- Low physical activity levels (**parf**)
- Inadequate Diet (**dietrf**)
- High consumption of alcohol (**alcoholrf**)

In the original data set the variable **alcoholpf** represented a protective factor and not a risk factor. Since we are interested in risk factors, that variable was transformed to represent a risk factor and it was called **alcoholrf**.

Once again, the confounding variables **sex** and **age** must be present in the model, as well as the variables related to inequalities, **education** and **ediscore**. The variable **region** was also added and will be used to make comparisons. The results of the model are summarized in Table 4.20 and the expression for the model obtained is the following:

$$E(Y) = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 sex + \hat{\beta}_3 education + \hat{\beta}_4 region + \hat{\beta}_5 ediscore\right) \qquad (4.8)$$

Table 4.20: Results from the Poisson regression model

	Estimate	**p-value**
intercept	0.827	<0.001
age	0.015	<0.001
sex		
female	-0.121	<0.001
education		
high school	-0.073	0.031
university education	-0.230	<0.001
region		
North	-0.046	0.412
Lisbon and Vale do Tejo	-0.111	0.131
Alentejo	-0.058	0.333
Algarve	-0.194	0.022
ediscore	0.017	0.093

When looking at the p-values it is possible to conclude that the variables age, sex, and education have significant effect on the number of risk factors, since their p-values are inferior to 0.05.

The coefficient for **age** is 0.015. This means that the expected log count for a one-unit increase in age is 0.015. In other words, with every unit increase, this predictor variable has multiplicative effect of 1.015 on the mean of the response variable, when controlling for other variables.

The estimate for **sex** is -0.121 and its exponential is 0.886. This shows that changing the sex category from male to female results in a decrease in the number of risk factors 0.886 times the intercept, because the estimate is negative. When changing the **sex** category from male to female the number of risk factors will decrease by 11.4% (1-0.886), when controlling for all other variables.

The estimate for the variable **education** in category "high school" is -0.073 and its exponential is 0.930. This shows that changing the **education** category from "no qualifications/pre-high school" to "high school" results in a decrease in the number of risk factors 0.930 times the intercept, because the estimate is negative. This means that the number of risk factors will decrease by 7% (1-0.930), when controlling for all other variables. When changing the **education** category from "no qualifications/pre-high school" to "university education" there will be a reduction of 20.5% in the number of risk factors.

The estimate for the variable **region** in category "North" is -0.046 and its exponential is 0.955. This shows that changing the **region** category from "Center" to "North" results in a decrease in the number of risk factors 0.955 times the intercept, because the estimate is negative. This means that the number of risk factors will decrease by 4.5% (1-0.955), when controlling for all other variables. When changing the **region** category from "Center" to "Lisbon and Vale do Tejo" there will be a reduction of 10.5% in the number of risk factors. When changing the **region** category from "Center" to "Alentejo" there will be a reduction of 5.6% in the number of risk factors. When changing the **region** category from "Center" to "Algarve" there will be a reduction of 17.7% in the number of risk factors. The difference

that is statistically significant is the one between Center and Algarve, since it is the only with a p-value inferior to 0.05.

The coefficient for **ediscore** is 0.017. This means that the expected log count for a one-unit increase in the ediscore is 0.017. In other words, with every one unit increase, this predictor variable has multiplicative effect of 1.017 on the mean of the response variable, when controlling for other variables.

4.4.1 Multiple Comparisons

The sampling option chosen in the e_COR study, makes it hard to draw solid conclusions when trying to make comparisons between regions or between age groups. This is because the sampling methodology does not reflect the way the Portuguese population is stratified, i.e., the samples were not weighted by the population size in each region making comparisons between regions harder to perform. Therefore, all the comparisons that are possible to make are within each region. The goal is to evaluate whether different levels of education have an impact on the number of risk factors an individual presents in each region.

The different levels of education were compared within each region and the results are summarized in Table 4.21, where education1 means "None/pre High School", education2 means "High School" and education3 means "University Education". The most significant difference (p-value <0.0001) was found between the highest and the lowest levels of education, respectively, University Education and None/pre High School education. The difference between High School and University Education was also significant (p-value of 0.0003). These results were the same in all regions.

Table 4.21: Results from *post hoc* multiple comparisons within each region

	Estimate	**p-value**
Region 1: North		
education1 – education 2	0.073	0.078
education1 – education 3	0.230	<0.0001
education2 – education 3	0.157	<0.001
Region 2: Center		
education1 – education 2	0.073	0.078
education1 – education 3	0.230	<0.0001
education2 – education 3	0.157	<0.001
Region 3: Lisbon and Vale do Tejo		
education1 – education 2	0.073	0.078
education1 – education 3	0.230	<0.0001
education2 – education 3	0.157	<0.001
Region 4: Alentejo		
education1 – education 2	0.073	0.078
education1 – education 3	0.230	<0.0001
education2 – education 3	0.157	<0.001
Region 5: Algarve		
education1 – education 2	0.073	0.078
education1 – education 3	0.230	<0.0001
education2 – education 3	0.157	<0.001

The results can be easily interpreted when observing Figure 4.16 where the blue bars are confidence intervals for the estimated marginal means and the red arrows are for the comparisons among them. If an arrow from one group overlaps an arrow from another group, the difference is not significant.

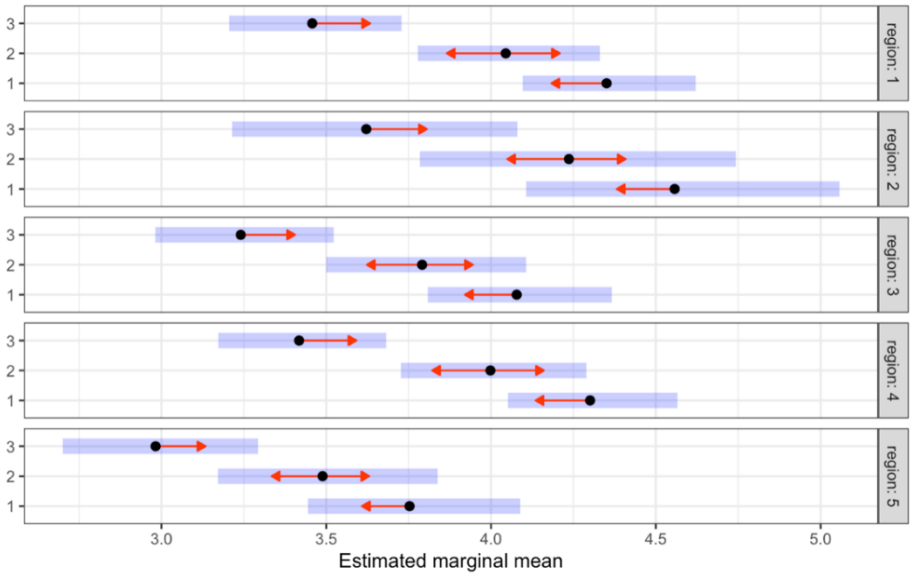

Figure 4.16: Estimated marginal means with respective 95% confidence intervals for the number of risk factors by education level and region (region 1-5 corresponds to North, Center, Lisbon and Vale do Tejo, Alentejo and Algarve)

Chapter 5 – Discussion and Results

The main goal of this project was to evaluate the possible effect of health inequalities, in particular the effect of low education and deprivation levels, in cerebro-cardiovascular disease, more specifically on the past occurrence of disease and number of risk factors presented by the individuals.

This work was conducted using data from a pre-existing study on the prevalence of cardiovascular risk factors of the Portuguese population, called e_COR, as well as data from INE and the Portuguese EDI. The e_COR database was very rich, containing demographic information about each participant, such as sex, age, region and municipality of residence, as well as lifestyle habits, metabolic parameters, physical characteristics, and medical history.

It is important to highlight the fact that the cerebro-cardiovascular events recorded happened in the past and the risk factors happen in the present. This means that the interpretation of the results is not the most conventional and that some of the variables would take different values if evaluated before the event happened in the individuals that are cases.

The first response variable considered was **cvevent**, representing the past occurrence of a cerebro-cardiovascular event. Modelling this variable allows the distinction between individuals who have had a cerebro-cardiovascular event in the past from those who have not. Logistic regression was the method used to model this variable, beginning with the confounding factors, sex and age, and selecting the relevant biological predictors through stepwise selection and, finally, adding the variables related to inequalities. The final model included the predictors **sex**, **age**, **cholmed**, **historyrf**, **smokenr**, **ldl**, **hypertensorsmed**, **education**, **ediscore** and the random effect of the **region** (4.4). Through the interpretation of this model, it was possible to conclude that the variables related to inequalities, **education** and **ediscore** are not considered statistically significant, since their p-values are far superior to the significance level of 0.05. In the presence of the other variables, knowing an individual's academic qualifications or level of deprivation is not relevant, because the other variables are better indicators of the past occurrence of a cerebro-cardiovascular disease. It was also possible to conclude that the random effect of the region is not statistically significant.

However, these results might not be accurate due to the data set being highly imbalanced, with only 84 individuals in a total of 1688 having had a cerebro-cardiovascular event. To try to overcome this problem, the SMOTE technique was performed to obtain a balanced data set. The final model determined for the balanced data set included the following predictors: **sex**, **age**, **cholmed**, **historyrf**, **waist**, **smokenr**, **ldl**, **tg*trigmed**, **beerunits**, **glucose*diabetesmed**, **dieteq**, **mealsday**, **vpa**, **hypertensorsmed**, **wineunits**, **alcoholpf**, **education**, **ediscore** and the random effect of the **region** (4.7). Through the interpretation of this model, it was possible to conclude that, once again, the variables related to inequalities, **education** and **ediscore** are not considered statistically significant, since their p-values are far superior to the significance level of 0.05. It was also possible to conclude that the random effect of the region is not statistically significant in this case as well, since all the 95% CI for conditional modes of the random effects overlap each other.

Both logistic regression models obtained suggest that when we have two individuals with similar physical and biological characteristics the influence of their education levels, deprivation levels and region of residence will not be a good indicator of whether they suffered a cerebro-cardiovascular event.

Through model validation techniques, such as determining the ROC curve and respective AUC as well as performing cross validation, both generalized linear mixed models determined showed high performance with elevated accuracy and low test errors.

The other response variable of interest was **nrfactors**, representing the number of risk factors each individual that participated in the e_COR study presented. This variable was modelled through Poisson regression obtaining a model with the covariables **age**, **sex**, **education**, **region** and **ediscore** (4.8). The goal of this was to evaluate whether different levels of education have an impact on the

number of risk factors an individual presents in each region. It was found a significant difference in the expected number of risk factors between individuals that completed university education and those that had none/pre-high school qualifications. The difference between individuals that completed high school and those that completed university education is also statistically significant. The variable **ediscore** had a p-value superior to the significance level of 0.05 indicating that the deprivation level of the municipality an individual lives in does not have a statistically significant effect on the number of risk factors to cerebro-cardiovascular disease the individual presents.

Previous studies have found that increased knowledge about risk factors for stroke is linked to higher socioeconomic position, meaning that people with a lower socioeconomic position (with regards to education, income, and occupation) are at a higher risk of having a stroke and have worse clinical outcomes compared to the general population. [29] In this study, the EDI was used as a socioeconomic indicator, however it seems not to be informative enough of the individual socioeconomic status since it did not demonstrate to have the importance as other studies reveal socioeconomic position has. This might be due to the fact that, in this study, the EDI is a socioeconomic indicator at the level of the municipality but in the same municipality there are people with very different individual socioeconomic positions. For future studies, it would be useful to consider an indicator at another level, the neighbourhood, for example.

The variables that characterize the access to healthcare available to the study participants were included in this work in addition to the individuals' characteristics. The fact that it was not found statistical relevance for these variables in this work does not mean they are not relevant, they are still important. Not finding differences related to levels of deprivation or region of residence, might reveal that the most important differences are in the individuals themselves.

There were other approaches available to model this data. Negative binomial regression could have also been used to model the response variable **nrfactors**, since it is an appropriate regression model for count data, as well as zero-inflated count models. For the response variable **cvevent**, other link functions could be used in the regression model.

In terms of the data, it would be interesting to collect data also in the autonomous regions of Madeira and Azores and not just mainland Portugal. The sample sizes should also be weighted by the population size in each area to obtain samples that are representative of the population.

It would also be interesting to conduct a follow-up study on the same sample to observe whether there were changes, either in the number of risk factors or in terms of cerebro-cardiovascular events.

References

[1] Global Health Europe. (2009, August 24). *Inequity and inequality in health*. Global Health Europe. https://globalhealtheurope.org/values/inequity-and-inequality-in-health/

[2] Arcaya, M. C., Arcaya, A. L., & Subramanian, S. V. (2015). Inequalities in health: definitions, concepts, and theories. *Global Health Action, 8*(1), 27106. https://doi.org/10.3402/gha.v8.27106

[3] WHO. (2018, February 22). *Health inequities and their causes*. Www.who.int. https://www.who.int/news-room/facts-in-pictures/detail/health-inequities-and-their-causes

[4] World Health Organization. "Cardiovascular Diseases (CVDs)." Who.int, World Health Organization: WHO, 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[5] Mayo Clinic. (2020). *Stroke - Symptoms and causes*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

[6] Sullivan, D. (2012). *Acute Myocardial Infarction: Causes, Symptoms, and Treatment*. Healthline. https://www.healthline.com/health/acute-myocardial-infarction

[7] CDC. (2020, September 8). *Peripheral Arterial Disease (PAD) | cdc.gov*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/PAD.htm

[8] *Cardiovascular diseases statistics Statistics Explained*. (2021). https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/37359.pdf

[9] INE. (2021, March 1). *Statistics Portugal - Web Portal*. Www.ine.pt. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=458514604&DESTAQUESmodo=2

[10] O'Donnell, M. J., Xavier, D., Liu, L., & INTERSTROKE investigators. (2010). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet (London, England), 376(9735), 112-23.* https://doi.org/10.1016/S0140-6736(10)60834-3

[11] Yusuf, S., Hawken, S., Ounpuu, S., & INTERHEART Study Investigators. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet (London, England), 364(9438), 937–952*. https://doi.org/10.1016/S0140-6736(04)17018-9~

[12] Ribeiro AI, Launay L, Guillaume E, Launoy G, Barros H (2018) The Portuguese version of the European Deprivation Index: Development and association with all-cause mortality. PLoS ONE 13 (12):e0208320.https://doi.org/10.1371/journal.pone.0208320

[13] Bourbon, M., Alves, A. C., & Rato, Q. (2019). Prevalência de fatores de risco cardiovascular na população portuguesa. In Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA, IP) (Ed.), https://www.insa.min-saude.pt/wp-content/uploads/2020/02/e_COR_relatorio.pdf

[14] Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, *24*(3), 69–71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

[15] Statistics Solutions. (2021). *Pearson's Correlation Coefficient*. Statistics Solutions. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/

[16] Glen, S. (2015). *Variance Inflation Factor*. Statistics How To. https://www.statisticshowto.com/variance-inflation-factor/

[17] Madsen, H., & Thyregod, P. (2011). *Introduction to General and Generalized Linear Models.* Chapman & Hall/CRCC. ISBN 978-1-4200-9155-7.

[18] McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models (2nd ed.). Boca Raton, FL:* Chapman and Hall/CRC. ISBN 0-412-31760-5.

[19] Turkman, M. A., & Loiola Silva, G. (2000). *Modelos Lineares Generalizados - da teoria à prática*. Sociedade Portuguesa de Estatística. https://www.spestatistica.pt/storage/app/uploads/public/5ff/0b8/26d/5ff0b826dacae404650097.pdf

[20] Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/bf03206482

[21] Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/jto.0b013e3181ec173d

[22] (2011) Leave-One-Out Cross-Validation. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_469

[23] Brownlee, J. (2018, May 21). *A Gentle Introduction to k-fold Cross-Validation*. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

[24] Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency*, *3*(2), 48–49. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4614595/

[25] McHugh, M. L. (2019). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[26] Delgado, R., & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PLOS ONE*, *14*(9), e0222916. https://doi.org/10.1371/journal.pone.0222916

[27] Yadav, D. (2020, April 14). *Weighted Logistic Regression for Imbalanced Dataset*. Medium. https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b

[28] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[29] Stack, K., Robertson, W. & Blackburn, C. Does socioeconomic position affect knowledge of the risk factors and warning signs of stroke in the WHO European region? A systematic literature review. BMC Public Health 20, 1473 (2020). https://doi.org/10.1186/s12889-020-09580-x

Appendices

Table 1: Detailed summary of stepwise selection process for original data

Step	Model Formula	AIC
1st	$-6.4873 - 0.5098\ \text{sexF} + 0.0516\ \text{Age} + 1.3999\ \text{cholmedY}$	558,79
2nd	$-6.9403 - 0.5929\ \text{sexF} + 0.0554\ \text{Age} + 1.4682\ \text{cholmedY} + 1.4619\ \text{historyrf1}$	542,24
3rd	$-7.4101 - 0.1984\ \text{sexF} + 0.0551\ \text{Age} + 1.4935\ \text{cholmedY} + 1.5249\ \text{historyrf1} + 0.0276\ \text{smokenr}$	529,66
4th	$-5.6881 - 0.17\ \text{sexF} + 0.0543\ \text{Age} + 1.2202\ \text{cholmedY} + 1.54\ \text{historyrf1} + 0.0283\ \text{smokenr} - 0.0137\ \text{ldl}$	519,18
5th	$-5.4686 - 0.1656\ \text{sexF} + 0.0431\ \text{Age} + 1.0930\ \text{cholmedY} + 1.5204\ \text{historyrf1} + 0.0274\ \text{smokenr} - 0.0130\ \text{ldl} + 0.7431\ \text{hypertensorsmedY}$	516,05
6th	$-5.4819 - 0.1726\ \text{sexF} + 0.0427\ \text{Age} + 1.1015\ \text{cholmedY} + 1.5140\ \text{historyrf1} + 0.0261\ \text{smokenr} - 0.0127\ \text{ldl} + 0.7186\ \text{hypertensorsmedY} + 0.6021\ \text{trigmedY}$	516,27
7th	$-6.3485 - 0.1866\ \text{sexF} + 0.0419\ \text{Age} + 1.0807\ \text{cholmedY} + 1.5184\ \text{historyrf1} + 0.0256\ \text{smokenr} - 0.0132\ \text{ldl} + 0.6324\ \text{hypertensorsmedY} + 0.5922\ \text{trigmedY} + 0.0368\ \text{bmi}$	516,48
8th	$-5.4819 - 0.1726\ \text{sexF} + 0.0427\ \text{Age} + 1.1015\ \text{cholmedY} + 1.5140\ \text{historyrf1} + 0.0261\ \text{smokenr} - 0.0127\ \text{ldl} + 0.7186\ \text{hypertensorsmedY} + 0.6021\ \text{trigmedY}$	516.27
9th	$-6.3485 - 0.1866\ \text{sexF} + 0.0419\ \text{Age} + 1.0807\ \text{cholmedY} + 1.5184\ \text{historyrf1} + 0.0256\ \text{smokenr} - 0.0132\ \text{ldl} + 0.6324\ \text{hypertensorsmedY}$	516,05

Table 2: Detailed summary of stepwise selection process for data obtained through SMOTE technique

Step	Model Formula	AIC
1st	$-6.1868 - 1.2359\ \text{sexF} + 0.0516\ \text{age} + 1.6099\ \text{cholmedY}$	694.38
2nd	$-7.2550 - 1.2456\ \text{sexF} + 0.0991\ \text{age} + 1.7270\ \text{cholmedY} + 2.1056\ \text{historyrfY}$	650.40
3rd	$-10.8978 - 1.0240\ \text{sexF} + 0.0803\ \text{age} + 1.7339\ \text{cholmedY} + 2.1197\ \text{historyrfY} + 0.0470\ \text{waist}$	622.86
4th	$-8.9530 - 0.9279\ \text{sexF} + 0.0725\ \text{age} + 1.4246\ \text{cholmedY} + 2.0374\ \text{historyrfY} + 0.0541\ \text{waist} - 0.0181\ \text{ldl}$	601.53
5th	$-9.1332 - 0.8372\ \text{sexF} + 0.0690\ \text{age} + 1.4175\ \text{cholmedY} + 2.1121\ \text{historyrfY} + 0.0571\ \text{waist} - 0.0196\ \text{ldl} + 1.4776\ \text{trigmedY}$	579.83
6th	$-9.1729 - 0.4959\ \text{sexF} + 0.0666\ \text{age} + 1.5018\ \text{cholmedY} + 2.1441\ \text{historyrfY} + 0.0524\ \text{waist} - 0.0182\ \text{ldl} + 1.3106\ \text{trigmedY} + 0.0294\ \text{smokenr}$	570.29
7th	$-9.3362 - 0.4887\ \text{sexF} + 0.0631\ \text{age} + 1.5503\ \text{cholmedY} + 2.1366\ \text{historyrfY} + 0.0575\ \text{waist} - 0.0179\ \text{ldl} + 3.8281\ \text{trigmedY} + 0.0332\ \text{smokenr} - 0.0176\ \text{tg} * \text{trigmedY}$	556.78

8th	$-9.1048 - 0.5802\ \text{sexF} + 0.0601\ \text{age} + 1.6193\ \text{cholmedY} + 2.1374\ \text{historyrfY} + 0.0572\ \text{waist} - 0.0174\ \text{ldl} + 3.8384\ \text{trigmedY} + 0.0342\ \text{smokenr} - 0.0171\ \text{tg} * \text{trigmedY} - 0.4947\ \text{beerunits}$	554.57
9th	$-10.0520 - 0.5257\ \text{sexF} + 0.0576\ \text{age} + 1.6346\ \text{cholmedY} + 2.1473\ \text{historyrfY} + 0.0569\ \text{waist} - 0.0172\ \text{ldl} + 3.7009\ \text{trigmedY} + 0.0369\ \text{smokenr} - 0.0161\ \text{tg} * \text{trigmedY} - 0.5351\ \text{beerunits} - 0.0201\ \text{glucose} * \text{diabetesmedY}$	555.79
10th	$-10.4387 - 0.5487\ \text{sexF} + 0.0546\ \text{age} + 1.6522\ \text{cholmedY} + 2.1283\ \text{historyrfY} + 0.0587\ \text{waist} - 0.0166\ \text{ldl} + 3.5883\ \text{trigmedY} + 0.0363\ \text{smokenr} - 0.0157\ \text{tg} * \text{trigmedY} - 0.5659\ \text{beerunits} - 0.0206\ \text{glucose} * \text{diabetesmedY} + 0.4975\ \text{dieteqY}$	553.47
11th	$-9.4702 - 0.4366\ \text{sexF} + 0.0528\ \text{age} + 1.6560\ \text{cholmedY} + 2.1946\ \text{historyrfY} + 0.0574\ \text{waist} - 0.0166\ \text{ldl} + 3.6516\ \text{trigmedY} + 0.0373\ \text{smokenr} - 0.0162\ \text{tg} * \text{trigmedY} - 0.6303\ \text{beerunits} - 0.0202\ \text{glucose} * \text{diabetesmedY} + 0.6248\ \text{dieteqY} - 0.2154\ \text{mealsday}$	551.08
12th	$-9.4702 - 0.4366\ \text{sexF} + 0.0528\ \text{age} + 1.6560\ \text{cholmedY} + 2.1946\ \text{historyrfY} + 0.0574\ \text{waist} - 0.0166\ \text{ldl} + 3.6516\ \text{trigmedY} + 0.0373\ \text{smokenr} - 0.0162\ \text{tg} * \text{trigmedY} - 0.6303\ \text{beerunits} - 0.0202\ \text{glucose} * \text{diabetesmedY} + 0.6248\ \text{dieteqY} - 0.2154\ \text{mealsday} + 2.4029\ \text{diabetesmedY}$	551.08
13th	$-8.9722 - 0.4936\ \text{sexF} + 0.0485\ \text{age} + 1.6468\ \text{cholmedY} + 2.1814\ \text{historyrfY} + 0.0558\ \text{waist} - 0.0167\ \text{ldl} + 3.6263\ \text{trigmedY} + 0.0380\ \text{smokenr} - 0.0166\ \text{tg} * \text{trigmedY} - 0.6556\ \text{beerunits} - 0.0206\ \text{glucose} * \text{diabetesmedY} + 0.6304\ \text{dieteqY} - 0.2126\ \text{mealsday} + 2.500\ \text{diabetesmedY} - 0.1748\ \text{vpa}$	548.97
14th	$-8.8456 - 0.4917\ \text{sexF} + 0.0421\ \text{age} + 1.5774\ \text{cholmedY} + 2.1249\ \text{historyrfY} + 0.0559\ \text{waist} - 0.0165\ \text{ldl} + 3.5705\ \text{trigmedY} + 0.0397\text{smokenr} - 0.0164\ \text{tg} * \text{trigmedY} - 0.6615\ \text{beerunits} - 0.0200\ \text{glucose} * \text{diabetesmedY} + 0.6355\ \text{dieteqY} - 0.1995\ \text{mealsday} + 2.3919\ \text{diabetesmedY} - 0.1850\ \text{vpa} + 0.4704\ \text{hypertesorsmedY}$	547.95
15th	$-7.8999 - 0.5389\ \text{sexF} + 0.0495\ \text{age} + 1.5798\ \text{cholmedY} + 2.1880\ \text{historyrfY} + 0.0551\ \text{waist} - 0.0160\ \text{ldl} + 3.3785\ \text{trigmedY} + 0.0393\ \text{smokenr} - 0.0153\ \text{tg} * \text{trigmedY} - 0.6422\ \text{beerunits} - 0.0202\ \text{glucose} * \text{diabetesmedY} + 0.6210\ \text{dieteqY} - 0.1874\ \text{mealsday} + 2.4054\ \text{diabetesmedY} - 0.1847\ \text{vpa} + 0.5247\ \text{hypertesorsmedY} - 0.0123\ \text{sbp}$	546.37
16th	$-7.8999 - 0.5389\ \text{sexF} + 0.0495\ \text{age} + 1.5798\ \text{cholmedY} + 2.1880\ \text{historyrfY} + 0.0551\ \text{waist} - 0.0160\ \text{ldl} + 3.3785\ \text{trigmedY} + 0.0393\ \text{smokenr} - 0.0153\ \text{tg} * \text{trigmedY} - 0.6422\ \text{beerunits} - 0.0202\ \text{glucose} * \text{diabetesmedY} + 0.6210\ \text{dieteqY} - 0.1874\ \text{mealsday} + 2.4054\ \text{diabetesmedY} - 0.1847\ \text{vpa} + 0.5247\ \text{hypertesorsmedY} - 0.0123\ \text{sbp} + 0.0131\ \text{glucose}$	546.37
17th	$-8.1139 - 0.6054\ \text{sexF} + 0.0550\ \text{age} + 1.6253\ \text{cholmedY} + 2.2135\ \text{historyrfY} + 0.0553\ \text{waist} - 0.0153\ \text{ldl} + 3.3858\ \text{trigmedY} + 0.0392\ \text{smokenr} - 0.0153\ \text{tg} * \text{trigmedY} - 0.6232\ \text{beerunits} - 0.0201\ \text{glucose} * \text{diabetesmedY} + 0.5981\ \text{dieteqY} - 0.2052\ \text{mealsday} + 2.4069\ \text{diabetesmedY} - 0.1989\ \text{vpa} + 0.5240\ \text{hypertesorsmedY} - 0.0128\ \text{sbp} + 0.0138\ \text{glucose} + 0.0679\ \text{mpa}$	545.65

18th	$-8.0093 - 0.7198\ \text{sexF} + 0.0493\ \text{age} + 1.6430\ \text{cholmedY} + 2.1593\ \text{historyrfY} + 0.0546\ \text{waist} - 0.0153\ \text{ldl} + 3.3887\ \text{trigmedY} + 0.0405\ \text{smokenr} - 0.0157\ \text{tg} * \text{trigmedY} - 0.5585\ \text{beerunits} - 0.0197\ \text{glucose} * \text{diabetesmedY} + 0.6209\ \text{dieteqY} - 0.2205\ \text{mealsday} + 2.3884\ \text{diabetesmedY} - 0.1944\ \text{vpa} + 0.5556\ \text{hypertesorsmedY} - 0.0107\ \text{sbp} + 0.0132\ \text{glucose} + 0.0739\ \text{mpa} - 0.1695\ \text{winunits}$	545.51
19th	$-6.9747 - 0.7929\ \text{sexF} + 0.0492\ \text{age} + 1.5950\ \text{cholmedY} + 2.1714\ \text{historyrfY} + 0.0542\ \text{waist} - 0.0157\ \text{ldl} + 3.6574\ \text{trigmedY} + 0.0395\ \text{smokenr} - 0.0170\ \text{tg} * \text{trigmedY} - 0.7391\ \text{beerunits} - 0.0205\ \text{glucose} * \text{diabetesmedY} + 0.5743\ \text{dieteqY} - 0.2288\ \text{mealsday} + 2.5071\ \text{diabetesmedY} - 0.2073\ \text{vpa} + 0.5961\ \text{hypertesorsmedY} - 0.0116\ \text{sbp} + 0.0150\ \text{glucose} + 0.0670\ \text{mpa} - 0.3721\ \text{wineunits} - 0.8978\ \text{alcoholpfY}$	540.44
20th	$-7.7083 - 0.8380\ \text{sexF} + 0.0518\ \text{age} + 1.6300\ \text{cholmedY} + 2.1205\ \text{historyrfY} + 0.0531\ \text{waist} - 0.0164\ \text{ldl} + 3.5448\ \text{trigmedY} + 0.0391\ \text{smokenr} - 0.0161\ \text{tg} * \text{trigmedY} - 0.7995\ \text{beerunits} - 0.0195\ \text{glucose} * \text{diabetesmedY} + 0.5705\ \text{dieteqY} - 0.2152\ \text{mealsday} + 2.4276\ \text{diabetesmedY} - 0.2100\ \text{vpa} + 0.5707\ \text{hypertesorsmedY} - 0.0194\ \text{sbp} + 0.0148\ \text{glucose} + 0.0637\ \text{mpa} - 0.3681\ \text{wineunits} - 0.9180\ \text{alcoholpfY} + 0.0226\ \text{dbp}$	540.57
21st	$-6.9747 - 0.7929\ \text{sexF} + 0.0492\ \text{age} + 1.5950\ \text{cholmedY} + 2.1714\ \text{historyrfY} + 0.0542\ \text{waist} - 0.0157\ \text{ldl} + 3.6574\ \text{trigmedY} + 0.0395\ \text{smokenr} - 0.0170\ \text{tg} * \text{trigmedY} - 0.7391\ \text{beerunits} - 0.0205\ \text{glucose} * \text{diabetesmedY} + 0.5743\ \text{dieteqY} - 0.2288\ \text{mealsday} + 2.5071\ \text{diabetesmedY} - 0.2073\ \text{vpa} + 0.5961\ \text{hypertesorsmedY} - 0.0116\ \text{sbp} + 0.0150\ \text{glucose} + 0.0670\ \text{mpa} - 0.3721\ \text{wineunits} - 0.8978\ \text{alcoholpfY}$	540.44
22nd	$-6.7290 - 0.7180\ \text{sexF} + 0.0489\ \text{age} + 1.5482\ \text{cholmedY} + 2.1538\ \text{historyrfY} + 0.0542\ \text{waist} - 0.0164\ \text{ldl} + 3.6689\ \text{trigmedY} + 0.0393\ \text{smokenr} - 0.0170\ \text{tg} * \text{trigmedY} - 0.7777\ \text{beerunits} - 0.0208\ \text{glucose} * \text{diabetesmedY} + 0.5930\ \text{dieteqY} - 0.2105\ \text{mealsday} + 2.5181\ \text{diabetesmedY} - 0.1953\ \text{vpa} + 0.5931\ \text{hypertesorsmedY} - 0.0113\ \text{sbp} + 0.0144\ \text{glucose} - 0.3625\ \text{wineunits} - 0.9279\ \text{alcoholpfY}$	541.00
23rd	$-7.5783 - 0.7032\ \text{sexF} + 0.0423\ \text{age} + 1.5504\ \text{cholmedY} + 2.0886\ \text{historyrfY} + 0.0547\ \text{waist} - 0.0169\ \text{ldl} + 3.8248\ \text{trigmedY} + 0.0400\ \text{smokenr} - 0.0181\ \text{tg} * \text{trigmedY} - 0.7727\ \text{beerunits} - 0.0203\ \text{glucose} * \text{diabetesmedY} + 0.6133\ \text{dieteqY} - 0.2230\ \text{mealsday} + 2.4742\ \text{diabetesmedY} - 0.1940\ \text{vpa} + 0.5508\ \text{hypertesorsmedY} + 0.0129\ \text{glucose} - 0.3989\ \text{wineunits} - 0.8992\ \text{alcoholpfY}$	541.85
24th	$-7.5783 - 0.7032\ \text{sexF} + 0.0423\ \text{age} + 1.5504\ \text{cholmedY} + 2.0886\ \text{historyrfY} + 0.0547\ \text{waist} - 0.0169\ \text{ldl} + 3.8248\ \text{trigmedY} + 0.0400\ \text{smokenr} - 0.0181\ \text{tg} * \text{trigmedY} - 0.7727\ \text{beerunits} - 0.0203\ \text{glucose} * \text{diabetesmedY} + 0.6133\ \text{dieteqY} - 0.2230\ \text{mealsday} + 2.4742\ \text{diabetesmedY} - 0.1940\ \text{vpa} + 0.5508\ \text{hypertesorsmedY} - 0.3989\ \text{wineunits} - 0.8992\ \text{alcoholpfY}$	541.85