Data Informed Decision Making

Cardiovascular Disease Prevention

Mohammad Ammar Faiq

Master of Science Thesis

Engineering and Policy Analysis
Faculty of Technology, Policy and Management
Delft University of Technology

TU Delft

Data Informed Decision Making

Cardiovascular Disease Prevention

by

Mohammad Ammar Faiq

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday August 19, 2019 at 13:00 AM.

Student number:	4697537	
Project duration:	February 1, 2019 – August 19, 2019	
Thesis committee:	Dr. S. W. Cunningham,	TU Delft, chair and supervisor
	Dr. H. G. van der Voort,	TU Delft, supervisor
	Dr. R. H. H. Groenwold,	Leiden University of Medical Center, Advisor
	drs. J. Kist M.D.,	Leiden University of Medical Center, Advisor

This thesis is confidential and cannot be made public until August 31, 2019.

TU Delft

Preface

This thesis is just some random things that I enjoyed so much. Venturing into the unknown world of micro-datasets has been really enjoyable and fun experience.

Special thanks to both of my parents and my family for supporting me with all their might and giving me hope as well as prayer during my studies from a start to the end.

Special thanks also to my thesis graduation committee : Scott Cunningham for all the brilliance ideas and guidance during my data adventure and thesis; Haiko van der Voort, with his upbringing and really nice suggestions on the relevance social context and big story line of my data exploration; Janet Kist as my day to day advisor at LUMC which gives outstanding support in practicalities of general practitioners into my research; Rolf Groenwold for guidance about his expertise in statistical analysis in medical research; and People from LUMC and EPA that support me during my thesis.

And lastly, I thank and congratulate to myself to be able to motivate yourself to finish your master degree with this amazing master thesis.

All the best to you Ammar in the future, I hope you get what you deserved in the end.

Kind regards,

Thesis …with a sense of …humor …and …shrouded …piece …of excitement …
Ha ha ha ha ha ha ha ……

Mohammad Ammar Faiq
Delft, August 2019

Contents

List of Figures

List of Tables

Executive Summary

Cardiovascular diseases are considered as one the deadliest disease and have also been the most prominent health burden around the world and particularly in the Netherlands. Enormously mitigation has been done to reduce the death burden by improving the quality of health care services and research related to cardiovascular diseases. One prominent strategy to reduce it is to identify early symptoms of cardiovascular diseases among the potential population. Currently, the prevailing cardiovascular disease risk prediction guidelines that used by a general practitioner only taking into account straightforward factors into their risk factors, and significant improvement to the guidelines is needed to include more socioeconomic factors into account since many expert realize the fallacy of the systems. This research proposes to expand the current quantitative model of cardiovascular risk estimation by including socioeconomic status variables. Thus, the following main research questions are proposed :

How are different socioeconomic factors contribute to cardiovascular diseases risk of the Hague population?

The cross-industry standard process for data mining (CRISP-DM) is used to structure the analysis as a way to be transparent, elaborate, and efficient in terms of how the modeling process is conducted. Firstly, business understanding is performed to gain several insights about all information related to cardiovascular diseases (clinical knowledge of the system of interest) and the quantitative analysis that is used to model the risk. Understanding these two aspects is vital to decide which variables and modeling approach should be taken into account in for the analysis. Secondly, data understanding are used to describe two things, how the data is produced, and what is the data looks like. The former one related to the data gathering process while the latter try to explore the data quality. During the data gathering process, information about how medical measurement code, medication code (ATC), death code (ICD), and diagnosis code (ICPC) are encoded in the dataset. In addition to that, data quality is check by defining two criteria, first is by using frequency count of all patients, all unique patients and all patients that experienced the first cardiovascular events; second is by defining the data sparsity of two datasets that are used (ELAN and CBS dataset).

Thirdly, data preprocessing are conducted to prepare data input for the modeling process. The first phase is data cleaning that tries to remove duplicated values and some unidentified and unnecessary rows from the dataset. Rows that do not have unique citizen id, have age below 25 years old and have data collection date below 2011 are removed from the dataset. Then the two datasets that are merged (ELAN and CBS) are imputed for tackling problems of the missing values (around 5 percent for CBS dataset and 80 percent missing values for ELAN dataset). Four methods of data imputation are performed, complete cases analysis, missing indicator method, single value imputation with mode, and multiple imputations. In addition to that, the data discretization is also performed to separate each variable into its categorical value. Thus, the data that is used in the modeling process consists of different socioeconomic variables and medical measurements from 2011 to 2017 with people from age 25 and above. The dependent variable that is chosen is ”First cardiovascular disease event,” which are either the first diagnosis or death by cardiovascular diseases.

Fourthly, the modeling process begins by performing Kaplan Meier analysis for different time axis, namely, follow up time, age at the baseline (from 2011), and dynamic age. The dynamic age has shown to be more informative compared with the other time axis given the context of the research; thus, it is chosen for this research. Then, the different parametric model is tested, and it is proven that Gompertz parametric model works best in estimating the survival rates of the dataset. After that, the Cox proportional hazards model are performed and set as one of the primary models that are chosen for this research. In order to filter out some of the socioeconomic variables, univariate and multivariate analysis are performed

with four different datasets that used different imputation methods. Backward elimination is performed for both analyses, which able to reduce ten socioeconomic variables from the equation. Then, all of the significant socioeconomic status variables are combined with medical measurements to performed Cox proportional hazards multivariate analysis. After that, Discrete Bayesian Network is performed from the dataset that has been discretized in the previous process. The first phase of the modeling began by defining the model structure. This step is performed by consulting different literature about cardiovascular disease and expert consultation for dependencies between different variables. After that, parameter learning is performed to define the multinomial distribution of the dataset given the model structure that is defined. Two approaches are used, maximum likelihood estimation, and Bayesian approach. Finally, the inference using Monte Carlo simulation with one million samples are used to syntactically simulated the dataset and probability table are used to shows the hazards ratio of individual variables (first cardiovascular events against the total number of a specific group of the population).

Then, in the fifth step, the result from the Cox proportional hazards model and Discrete Bayesian Network model are validated and evaluated. Expert validation is used in both of the models, while also Schoenfeld residuals tests are used in the Cox proportional hazards model to check the possibility of violation in proportional hazards assumption. On the other hand, Discrete Bayesian Network is validated by comparing the survival curve of the Cox proportional hazards model with age as a time axis, with the ”simulated” survival curve in Discrete Bayesian Network. However, although the model is considered to be validated correctly, there is certain doubt about the precise number and coefficients that are produced by both of the models. Thus, to avoid misinterpretation and overestimating the risk of a particular group of people, only comparative risk order between different groups are interpreted as the modeling result that is concluded.

It is without a doubt that older people have a higher chance of having cardiovascular diseases. In this research, the Cox proportional hazards model shows that the population starts to experience the first cardiovascular events from age 45 years old forward with the median (50 percent of the population have an event or censored) reached at the age of 75 years old. While being a male also increase the chance of having cardiovascular disease compared to female. Besides, people that have higher income, higher prosperity (income and wealth or assets) and live above the poverty line have a good prognostics compared to their counterparts (the lower values). Good prognostics means that they have a lower chance of having the first cardiovascular disease, thus lowering their risk of having cardiovascular disease. Additionally, the group of a single man in pension age (above 67 years old) have a better prognostics compared to a single man that is not from pension age. Meanwhile, people that have income from property have a lower risk of having cardiovascular diseases compared to people that have income from salary.

The same thing can be observed in the Discrete Bayesian Network, with people that have better financial status have a lower risk of having the first cardiovascular diseases compared to the one with worse financial status. Meanwhile, it is observed that some ethnic minority group are most susceptible such as Turkish, Moroccans, Antillean and Aruba, and especially Polish people. The result regarding Polish people that considered to be the most susceptible groups among all other ethnicities confirm the beliefs and norms of general practitioners that are not covered in the NHG guidelines for general practitioners in the Netherlands for cardiovascular diseases prevention act.

Despite the points mentioned above, when considering to include socioeconomic status in the cardiovascular risk prediction model, one should not overestimate the result that comes from the model. Thus, the use of socioeconomic status variables is therefore primarily as means for the practitioner to encourage patient to take the medical measurement which then can be used as the primary indicator to make the judgment for cardiovascular diseases prevention.

Therefore, it is recommended the focus of cardiovascular diseases prevention program should be in the group of people with age more than 50 years old that live below the poverty line (poor financial condition) and are part of ethnic minority group especially, Turkish, Moroccans, Antillean and Aruba, and Polish people.

1

Introduction

Cardiovascular diseases were the most significant burden of disease in the Netherlands, which account for 51,266 deaths or 35 percent of total mortality in 1997 (CBS, 2018a). The Dutch government has made a various attempt for reducing the risk of death by cardiovascular diseases by increasing the budget for health care services (OECD, 2018). One of the most significant expenditure included acute care for cardiovascular disease patient and cardiovascular research. It was proven to be effective since the mortality dropped significantly, which now death caused by cardiovascular disease was responsible for 38,199 deaths or roughly 12.5 percent of total death in 2017 (CBS, 2018a). This fact is also considered as lower than average in European Union (OECD, 2018).

Despite Netherlands significant improvements in mitigating cardiovascular disease, the problem associated with it persist, especially in providing better and the right treatment for the most vulnerable individual. One of the issues is related to early identification of cardiovascular disease symptoms that can occur among the population. Early identification of symptoms related to cardiovascular disease could help significantly to prevent the undesired condition that could happen in the future (Piepoli et al., 2016). Currently, each general practitioner used generic guidelines (Figure 1.1) for suggesting a potential cardiovascular disease patient go through a particular procedure. The factors that are considered for these guidelines used some standard medical measurements, such as, blood pressure, blood count, smoking, gender, cholesterol, diabetes, and age (Woodward et al., 2007). This generic guideline that used for deciding which measurement, treatment and medication sometimes failed to identify some of the relevant and vital cases of patients, especially in identifying the potential cardiovascular diseases risk among younger patients with different ethnic groups (Bos et al., 2004; Perini et al., 2018). Apparently, some experienced general practitioner already realizes the shortcoming of the current guidelines (SCORE, Figure 1.1), especially among ethnic minority group such as Polish or Surinamese people. As a result, it is not that uncommon that general practitioner suggests that some patients to undergo some medical measurement to confirm the potential cardiovascular even though that particular individual is not included as the patient that require medical inspection. However, for a younger general practitioner that used the guidelines as it is, they might not realize the shortcoming of the guidelines and therefore, can not perform the cardiovascular prevention efficiently. Thus, to aid the general practitioner in decision making for cardiovascular diseases prevention, more detailed guidelines are needed which incorporate not only medical measurement but also some socioeconomic status. As the first step for making more comprehensive guidelines that incorporate socioeconomic status, there is a need for more elaborated cardiovascular disease risk prediction model to be created prior to the guidelines. These research aim to explore the possibility of incorporating socioeconomic variables in different cardiovascular disease risk prediction model, which then serve as fundamental of making the new cardiovascular diseases prevention guidelines.

Bloeddruk	Leeftijd	Niet-roker						Roker						
180		7	8	10	12	15	18	13	15	18	21	26	31	Sterfte
		22-28	26-33	31-39	37-48	46-58	>50	40-51	47-60	>50	>50	>50	>50	Ziekte + Sterfte
160		5	6	7	9	11	13	9	11	13	16	19	23	Sterfte
	65	15-20	18-23	22-28	27-34	33-42	41-53	29-37	34-43	40-52	49-62	>50	>50	Ziekte + Sterfte
140		3	4	5	6	8	10	7	8	9	11	14	17	Sterfte
		11-14	13-17	16-20	19-25	24-30	30-38	20-26	24-31	29-37	35-45	44-56	>50	Ziekte + Sterfte
120		2	3	4	4	5	7	5	5	7	8	10	13	Sterfte
		8-10	9-12	11-14	14-18	17-22	22-28	14-18	17-22	21-27	25-32	32-40	39-50	Ziekte + Sterfte
180		4	5	6	8	10	12	8	10	12	15	18	22	Sterfte
		16-20	19-24	23-29	28-36	35-45	44-56	30-38	35-45	43-54	>50	>50	>50	Ziekte + Sterfte
160		3	4	5	6	7	9	6	7	9	11	13	16	Sterfte
	60	11-14	14-17	16-21	20-26	25-32	32-40	21-27	25-32	31-39	37-47	46-58	>50	Ziekte + Sterfte
140		2	3	3	4	5	6	4	5	6	8	9	12	Sterfte
		8-10	10-12	12-15	14-18	18-23	23-29	15-19	18-23	22-28	27-34	33-42	42-53	Ziekte + Sterfte
120		2	2	2	3	4	5	3	4	4	5	7	9	Sterfte
		6-7	7-9	8-11	10-13	13-16	16-21	11-13	13-16	16-20	19-24	24-30	30-38	Ziekte + Sterfte
180		3	3	4	5	6	8	5	6	8	9	12	15	Sterfte
		11-13	13-16	16-20	19-25	24-31	30-39	19-25	24-30	29-37	36-45	44-56	>50	Ziekte + Sterfte
160		2	2	3	4	5	6	4	4	5	7	8	11	Sterfte
	55	7-9	10-11	11-14	14-17	17-22	22-28	14-18	17-22	21-26	26-33	32-41	40-51	Ziekte + Sterfte
140		1	2	2	3	3	4	3	3	4	5	6	8	Sterfte
		5-7	6-8	8-10	10-12	12-16	16-20	10-13	12-15	15-19	18-23	23-29	29-37	Ziekte + Sterfte
120		1	1	1	2	2	3	2	2	3	3	4	5	Sterfte
		4-5	4-6	6-7	7-9	9-11	11-14	7-9	9-11	10-13	13-17	16-21	21-27	Ziekte + Sterfte
180		2	2	2	3	4	5	3	4	5	6	7	9	Sterfte
		6-8	8-10	10-12	12-15	15-19	20-25	12-16	15-19	18-23	23-29	28-36	36-46	Ziekte + Sterfte
160		1	1	2	2	3	4	2	3	3	4	5	6	Sterfte
	50	4-6	6-7	7-9	9-11	11-14	14-18	9-11	10-13	13-16	16-20	20-26	26-33	Ziekte + Sterfte
140		1	1	1	2	2	3	2	2	2	3	4	5	Sterfte
		3-4	4-5	5-6	6-8	8-10	10-13	6-8	7-9	9-12	11-15	15-18	19-24	Ziekte + Sterfte
120		1	1	1	1	1	2	1	1	2	2	3	3	Sterfte
		2-3	3-4	3-4	4-6	6-7	7-9	4-5	5-7	7-8	8-10	10-13	13-17	Ziekte + Sterfte
180		1	1	1	2	2	3	2	2	2	3	4	5	Sterfte
		4-5	5-6	6-7	7-9	9-11	12-15	7-9	9-11	11-13	13-17	17-21	22-27	Ziekte + Sterfte
160		1	1	1	1	2	2	1	1	2	2	3	4	Sterfte
	45	3-3	3-4	4-5	5-6	7-8	8-11	5-6	6-8	8-10	10-12	12-15	16-20	Ziekte + Sterfte
140		<1	1	1	1	1	1	1	1	1	2	2	3	Sterfte
		2-2	2-3	3-4	4-5	5-6	6-8	4-4	4-5	5-7	7-9	9-11	11-14	Ziekte + Sterfte
120		<1	<1	<1	1	1	1	1	1	1	1	1	1	Sterfte
		1-2	2-2	2-3	3-3	3-4	4-5	2-3	3-4	4-5	5-6	6-8	8-10	Ziekte + Sterfte
		3	4	5	6	7	8	3	4	5	6	7	8	
		TC-HDL-ratio						TC-HDL-ratio						

In de vakjes staat het tienjaarssterfterisico als gevolg van hart- en vaatziekten, evenals een indicatie van het risico op ziekte plus sterfte.

Figure 1.1: Example of SCORE chart for male population (Source: NHG (2019a)).

1.1. Research Scope

This research focuses on cardiovascular diseases prevention among the Hague population that also includes younger population. The event of interest that this research particularly

interested is ”The first cardiovascular diseases”, which could be either diagnosis or sudden death by cardiovascular diseases. In addition to that, the Hague is chosen in this research as a region of interests since it is considered as one the most segregated cities (Hartog and Zorlu, 2009) in the Netherlands, and could provide a better insight related to identifying cardiovascular disease-prone population cohorts that might exist in the Netherlands. The second reason why the Hague is chosen is because of the ELAN dataset that is used in this research are consists only medical records of a patient who live and registered in various general practitioners in the Hague. Therefore, as to limit the scope of analysis, the second dataset from CBS microdataset which consists various socioeconomic status for each individual across the Netherlands are going to be reduced according to the dataset in ELAN dataset.

1.2. Research Objective

As mentioned before, in general this research aims to aid general practitioner that is considered as a decision-maker to make a better judgment of patient treatment and improve knowledge transfer between general practitioner across the Netherlands through more elaborated guidelines by including not only medical measurements and physiological characteristics but also socioeconomic status (SES) that could be associated with cardiovascular disease risk. There are two main objectives for this research, exploring various socioeconomic status that could potentially be used in the cardiovascular risk model. And exploring the possibility of incorporating the socioeconomic status in the cardiovascular risk prediction model.

1.2.1. Socioeconomic status variable exploration

Thus to be able to study the various effect of socioeconomic status to the cardiovascular diseases, this research will try to utilize a theory that comes from social epidemiology studies branch. It investigates the influence of social distribution and social determinants on individual and population health. It also explores which socio-structural factors influence health and diseases. Socio-structural factors, such as ethnicity, social class, social network, social capital, income distribution, social policy; are often chosen as one factor that could determine the underlying reason why people got certain kind of diseases (Honjo, 2004).

In addition to that, there are specific socioeconomic characteristics that might exist among the population that can lead to a particular risk in cardiovascular diseases. These characteristics can be used as an indicator of the risk value among the population. In fact, categorizing factors that individual has then associate it with specific risk value in a patient has been done in some research, whether it is related to socioeconomic status (Goldstein et al., 2017b) or medical (biological) status (Bos et al., 2004). Since the Hague, in particular, is considered as highly segregated city with diverse ethnicity and socioeconomic status (Hartog and Zorlu, 2009; Perini et al., 2018), it is useful, before predicting the future risk of cardiovascular diseases risk in the population of the Hague, to first look into the past and current situation by associating the risk based on population characteristics.

1.2.2. Model exploration

Most cardiovascular disease risk studies used a Cox-proportional hazards model to determine the survival distribution. Studies such as Framingham (D’Agostino et al., 2008), QRISK (Hippisley-Cox et al., 2011), PROCAM (Assmann et al., 2002); are some of the examples of cardiovascular risk studies that used a Cox proportional hazards model. Besides that, the Weibull model is also used to explain the survival rate of patients to determine the risk of cardiovascular diseases, such as SCORE (Assmann et al., 2002) and ASSIGN (Assmann et al., 2002) project. There are also possibilities to use the Gompertz model to assess cardiovascular disease risk, as mentioned by Lee et al. (2014) and Beltrán-Sánchez et al. (2012). Lastly, opportunities also exist to combine each of those, therefore, producing Cox-Weibull or Cox-Gompertz model (Hall, 2016). Each of these models has their assumption; for example, Cox-proportional hazards assume that the ratio of the hazards for any two individuals is constant over time. On the other hand, the Weibull hazards model assumes that a specific patient rate of survival against cardiovascular diseases is following Weibull distribution.

The risk models that mentioned above used (frequency-based) statistical inference which most often the definition that is used is misunderstood by many medical scientists since it relies heavily on the context of the research and the sample size (population size) of the empirical data (Gurrin et al., 2000). Many works of the literature suggest that there are advantages of using the Bayesian approach for treatment recommendation in clinical practice to mitigate the drawbacks of (frequency-based) statistical inference (Nguefack-Tsague, 2011; van Gerven et al., 2008; Verduijn et al., 2007). One of the main advantages is the ability to dynamically change of what is known from the past clinical trials (prior) with new information from new clinical trials while also incorporating the uncertainties factors into the prediction (Bittl and He, 2017).

1.3. Research Structure

Formally, there are at least three well-known frameworks for data mining processes, namely KDD (Knowledge Discovery in Database)(Fayyad et al., 1996), SEMMA (Sample, Explore, Modify, Model, and Assess), and CRISP-DM (Cross-industry standard process for data mining). When compared, the literature suggests that CRISP-DM provide more concrete explanations of each phase of the data mining processes (Azevedo and Santos, 2008) than SEMMA and can act as an implementation of KDD processes. It is, therefore, beneficial for the current development of this research to use CRISP-DM as an initial framework for conducting the whole data mining phases. The cross-industry standard process for data mining (CRISP-DM) consists of business understanding, data understanding, data preprocessing, data modeling, evaluation, and deployment (Fayyad et al., 1996). The business understanding part is covered in the first three chapters, with the first and second chapter discuss the motivation of the research, as well as various variables that essential to consider in cardiovascular risk modeling. The third chapter will discuss the necessary information and theory of the data modeling that will be discussed in the fourth chapter. Data preprocessing and modeling will be addressed in the fourth chapter, while the evaluation of the model will be discussed in the fifth chapter. Lastly, the research will be concluded and discussed in the last chapter, which conclude the cycle of CRISP-DM.

1.4. Research Question

In the current situation, early literature research (will be discussed in Chapter 2 and 3) by the authors shows only few quantitative research that includes socioeconomic factors to measure cardiovascular risk in The Hague population. Therefore the authors consider this topic as a knowledge gap and will try to base the master thesis research on revolving around these issues. Finally, Considering knowledge gap between conventional cardiovascular diseases risk prediction scheme and the socioeconomic context of the population, this research will try to offer new possibilities for the use of time-series statistical modeling (Bittl and He, 2017; Gurrin et al., 2000), machine learning techniques (Akhil et al., 2012; Goldstein et al., 2017a,1), and modeling and simulation concept (Nianogo and Arah, 2015), to have risk prediction of cardiovascular diseases events with improved and added predictors. Therefore, the main and sub-research question of this master thesis will be:

How are different socioeconomic factors contribute to cardiovascular diseases risk of the Hague population?

Sub-research question:

- *"What are the socioeconomic status that is feasible to include in the cardiovascular risk model*?"

- *"What are the trade-off between the various model that is used for cardiovascular risk modelling*?"

2

Literature Review

Like many European countries, Netherlands healthcare systems influenced by the Bismarckian social welfare system that is gradually changing since 1860 with several policy reforms. The healthcare systems are shared among several actors, government, health professional, and insurers with extensive regulation that governs their interaction. Therefore, before going into the role of institution and regulation of Dutch health systems related to cardiovascular disease, it is useful to know the brief history of policy reform relate to it as well as the organization structure for the sake of knowing the how the decision was made for a patient related to cardiovascular disease

This literature review was conducted to gain more information about how the dutch healthcare system governs itself for cardiovascular prevention activities. Then, discussion about necessary variables to consider for the cardiovascular risk prediction model is discussed by considering various literature while also putting the GP decision making the process into perspective.

This literature review will consist of 4 different parts, starting from the historical description of health policy reform, until the latest development of cardiovascular disease research. The primary source of information of the first and the second part of this chapter mainly came from health system review (Exter et al., 2004; Schäfer et al., 2010) and several official website (Overheid, 2019a,1; Volksgezondheid, 2019c). This was done by following all of the policy reform and organizations keywords that mentioned in the report, such as the Health Act (*Gezondheidswet*), or *Gemeentelijk Gezondheidsdienten* (GGDs). Then, several governmental websites were visited to gain more detailed information about the subjects. Subsequently, the last two subsections will discuss how general practitioner in the Dutch healthcare systems are governing themselves and the implication of the decision making the process to the variables that used in the cardiovascular risk prediction model.

2.1. Dutch Healthcare Systems Policy Reforms

Healthcare system in the Netherlands is continuously changing with a lot of policy reform among different level. Although there are many reforms in the systems, the most significant changes that contribute to the current healthcare system can be separated into three major reformations. The first most significant reform date back in 1865 to early 1900 through ”Health Act” (*Gezondheidswet*), which lay the foundation of improving the health of a citizen and also define the role and involvement of government in improving the public health sector (Overheid, 2019c). Since then, through municipal health services (*Gemeentelijk Gezondheidsdienten*, GGDs) government has been partly involved in the public health issues (Overheid, 2019a). Additionally, around these reformations, the government also separated several inspection areas, namely healthcare, pharmaceutical care, mental healthcare, and veterinary care. Then, in 1995, the Health Care Inspectorate (*Inspectie Gezondheidszorg*, IGZ or now called health and youth care inspectorate *Inspectie Gezondheidszorg en Jeugd*, IGJ) was established which combined healthcare, pharmaceutical, and mental care and considered as

a separate body of Ministry of Health, Welfare and Sport (Volksgezondheid, 2019c).

The second most significant policy reform happened in 2006 with Health Insurance Act (*Zorgverzekeringswet*, Zvw) that introduced managed competition concept in the public health sector that also obligates the Dutch citizen to have a health insurance (Exter et al., 2004). This law gives more mandate to the health insurers, insured and providers to choose and accept health care package for their own. This law also places the role of government as a regulator and supervisor of the healthcare systems and market, which the three actors need to compete (see section 2.2 for more details). Therefore, the Health Insurance Act structured the Dutch healthcare system in such a decentralized manner with multiple actors that have a different active role in the system.

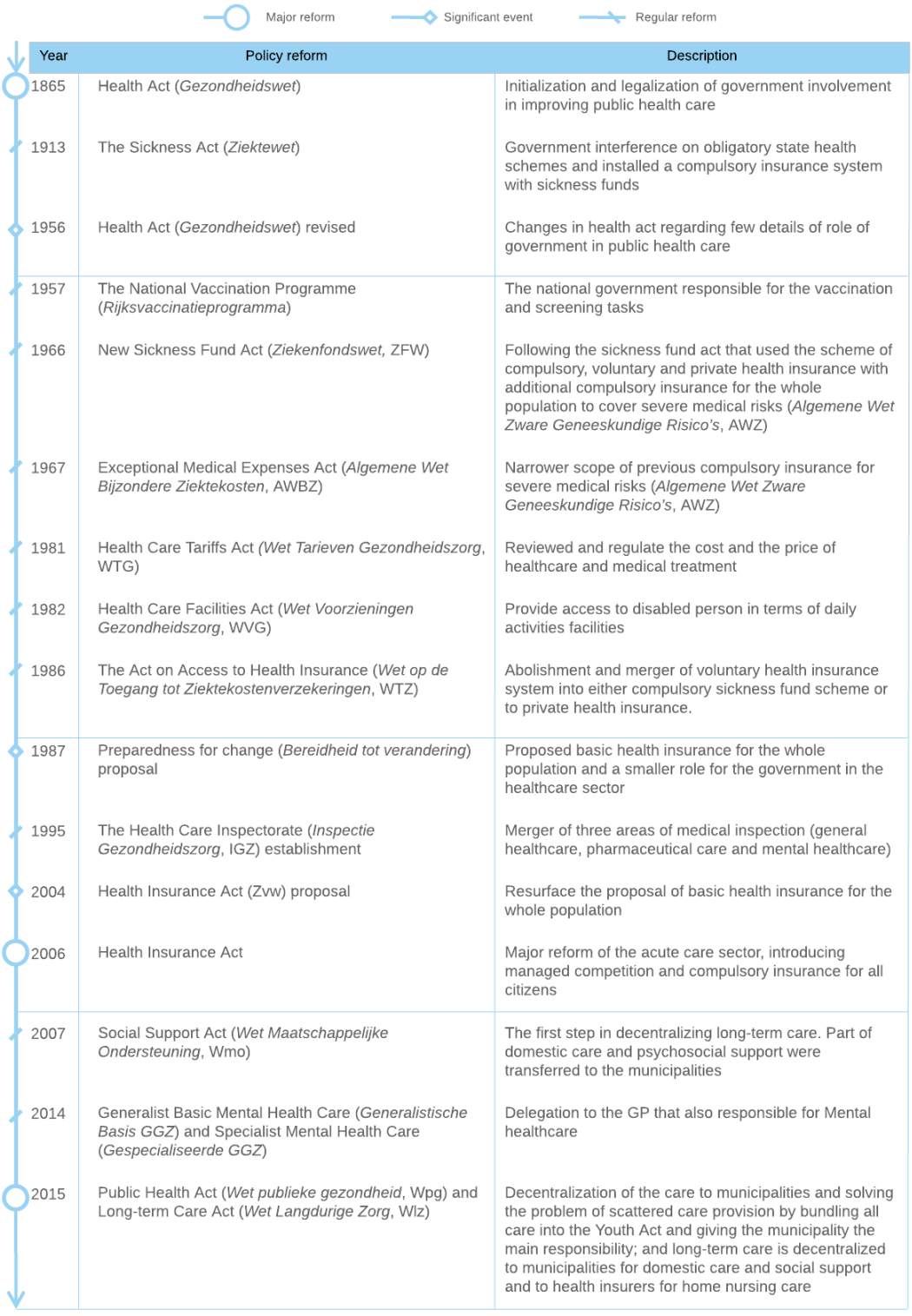
Lastly, in 2015, the Dutch healthcare system has been major reform and reorganized to be more decentralized through the Public Health Act (*Wet publieke gezondheid*, Wpg) (Overheid, 2019e). The reform started from the reallocation of the Exceptional Medical Expenses Act (AWBZ) (Overheid, 2019b); delegating personal care, home nursing, and long-term mental healthcare to the Health Insurance Act (Zvw) and support care, daycare and youth (mental health) care services to the Social Support Act (Overheid, 2019d); introduce the Long-term Care Act (*Wet Langdurige Zorg*, Wlz) for vulnerable elderly and disable people (Overheid, 2019f). In addition to all those reforms, Ministry of Health, Welfare, and Sport, put more power into municipalities in deciding and implementing the policy for the health sector as well as the budget for it (IHCSP, 2019). For more details about Dutch healthcare policy reform, please have a look at Figure 2.1.

2.2. Dutch Healthcare Organization Structure

Separate subsystems can be derived from the Dutch healthcare system based on the role of each care in the system. As a matter of fact, the Dutch systems are unique with much decentralization between the care that needs to be provided to the patient. When discussing health care systems in the Netherlands, people usually refer to the term ”curative care”. It means that medical care or practices that nurture the patient intending to cure them. In this subsystems, exist, three different actors, starting from health insurers, health providers, and patients or citizen with the different market for each of their interaction that is strictly monitored by the independent governmental bodies. Health insurers are responsible for deciding which health package that citizen have based on the material compensation that citizen gives to the insurers. They are also obliged to accept any person that wishes to get an insurance (Overheid, 2019d) without any exception for conditions. Besides that, they also responsible for negotiating the health package with the healthcare providers for the patients that they insure. The patients also play an active role in deciding which health insurers they are going to get the package and which package to choose. They also have freedom of choice regarding which health providers they want to get treatment. Lastly, there are health providers that cure or treat a patient with their expertise and facilities available. They also decided on the cost of the health services that they give to a patient. Although citizen is obliged to pay for the insurance, most of the curative care services budget still publicly funded (IHCSP, 2019).

On the other hand, long term care refers to treatments or service that targeted to an individual with some severe conditions that makes them unable to perform normal activities by themselves. Severe conditions include some mental problems, disabilities, or even some chronic diseases, which means they need support from medical and non-medical services. The Long-term Care Act (Overheid, 2019d) delegate municipalities for domestic services and youth care and insurers for nursing home and home nursing care with the support from the national government through multiple legislation (IHCSP, 2019). In the case of government, municipalities act as an executor of the program while the national government take a more distant role for the long term care. The rules and funding for long-term care are regulated not only in WIz but also Wpg and Wmo, which managed by the national government.

Several independent and dependent governmental institution exist to help keep the healthcare systems running as well as to improve the current system. Moreover, there is three independent institutions that quite crucial in regulating the systems, such as Consumers



Year	Policy reform	Description
1865	Health Act (*Gezondheidswet*)	Initialization and legalization of government involvement in improving public health care
1913	The Sickness Act (*Ziektewet*)	Government interference on obligatory state health schemes and installed a compulsory insurance system with sickness funds
1956	Health Act (*Gezondheidswet*) revised	Changes in health act regarding few details of role of government in public health care
1957	The National Vaccination Programme (*Rijksvaccinatieprogramma*)	The national government responsible for the vaccination and screening tasks
1966	New Sickness Fund Act (*Ziekenfondswet,* ZFW)	Following the sickness fund act that used the scheme of compulsory, voluntary and private health insurance with additional compulsory insurance for the whole population to cover severe medical risks (*Algemene Wet Zware Geneeskundige Risico's*, AWZ)
1967	Exceptional Medical Expenses Act (*Algemene Wet Bijzondere Ziektekosten*, AWBZ)	Narrower scope of previous compulsory insurance for severe medical risks (*Algemene Wet Zware Geneeskundige Risico's*, AWZ)
1981	Health Care Tariffs Act *(Wet Tarieven Gezondheidszorg*, WTG)	Reviewed and regulate the cost and the price of healthcare and medical treatment
1982	Health Care Facilities Act (*Wet Voorzieningen Gezondheidszorg*, WVG)	Provide access to disabled person in terms of daily activities facilities
1986	The Act on Access to Health Insurance (*Wet op de Toegang tot Ziektekostenverzekeringen*, WTZ)	Abolishment and merger of voluntary health insurance system into either compulsory sickness fund scheme or to private health insurance.
1987	Preparedness for change (*Bereidheid tot verandering*) proposal	Proposed basic health insurance for the whole population and a smaller role for the government in the healthcare sector
1995	The Health Care Inspectorate (*Inspectie Gezondheidszorg*, IGZ) establishment	Merger of three areas of medical inspection (general healthcare, pharmaceutical care and mental healthcare)
2004	Health Insurance Act (Zvw) proposal	Resurface the proposal of basic health insurance for the whole population
2006	Health Insurance Act	Major reform of the acute care sector, introducing managed competition and compulsory insurance for all citizens
2007	Social Support Act (*Wet Maatschappelijke Ondersteuning*, Wmo)	The first step in decentralizing long-term care. Part of domestic care and psychosocial support were transferred to the municipalities
2014	Generalist Basic Mental Health Care (*Generalistische Basis GGZ*) and Specialist Mental Health Care (*Gespecialiseerde GGZ*)	Delegation to the GP that also responsible for Mental healthcare
2015	Public Health Act (*Wet publieke gezondheid*, Wpg) and Long-term Care Act (*Wet Langdurige Zorg*, Wlz)	Decentralization of the care to municipalities and solving the problem of scattered care provision by bundling all care into the Youth Act and giving the municipality the main responsibility; and long-term care is decentralized to municipalities for domestic care and social support and to health insurers for home nursing care

Figure 2.1: Timeline of Healthcare reform in the Netherlands

and Market Authority (ACM), Dutch Healthcare authority and Health (NZa) and Health and Youth care Inspectorate (IGJ). ACM is an independent authority that is operating straight from Dutch law that is not under any ministry. In terms of the healthcare sector, they ensure that healthcare insurers fair and follow the rules in terms of the cost and the health package that the citizen receive while actively monitor that citizen (consumers) to be always informed and their rights protected (Markt, 2019). Then, there is NZa that regulate the market in which healthcare systems operates. They make sure that the citizen that entitled with the insurance by health insurers receive the health package quality that they promise. The same applies to health insurers, NZa makes sure that quality of health providers are up to the standard that the health insurers negotiate with the providers. In addition to that, NZa also plays a role in research and development of public health and healthcare systems to advise the Ministry of Health, Welfare, and Sport. Even though NZa is part of the ministry, they act independently and capable of making the decision themselves related to health market regulation if deemed necessary (Volksgezondheid, 2019a). Lastly, IGJ or also known as Health care inspectorate (IGZ) which regulate, supervise, and enforce healthcare providers to produce good quality of services and expertise (Volksgezondheid, 2019a). As previously mentioned in the timeline, IGZ was first established in 1995 and until just recently changed into IGJ in 2016 after the Youth Act in 2015 (Volksgezondheid, 2019d) to merge the IGZ with The Inspectorate for Youth Care (*Inspectie Jeugdzorg*, IJZ). Then the new established IGJ responsibility becoming to oversee not only the health providers quality but also youth care as well.

Moreover, there are certain institutions that in charge specifically for research and development in Health sector and advise strategically for the ministry of health, welfare and sport (see advisory bodies in figure 2.2). Based on the scope of research that each advisory body conduct and the type of advice that they give to the ministry, three types of institutions can be classified; general public health and society, well-being from the social, parliamentary and some technical aspect. Firstly, the Council for Public Health and Society (*Raad voor Volksgezondheid en Samenleving*, RVS) and The Social and Cultural Planning Office (*Sociaal en Cultureel Planbureau*, SCP) are the two organizations that handles mostly public health and society well-being policy (Volksgezondheid, 2019b) as well as the social and cultural well-being of the citizen in the Netherlands (SCP, 2019). Secondly, the parliamentary and legal issues of the health sector are handled by the Health Council (*Gezondheidsraad*). As stated and regulated by The Health Act, The Health council tasks are to provides requested and unrequested advice to government and parliament on issues across the public health spectrum: from health care, disease prevention and nutrition to living environments, working conditions and innovation, and knowledge infrastructure (Volksgezondheid, 2018). Lastly, there is the Dutch National Institute for Public Health and the Environment (*Rijksinstituut voor Volksgezondheid en Milieu*, RIVM) that handles most of the research in the health sector more detailed and technical. RIVM most often collaborate with research institutes and/or all university across the Netherlands to gain knowledge and improve public health and the environment from the end-to-end standpoint of view. Therefore, it is not strange that RIVM also coordinates screening programs that are necessary to mitigate some of the infectious or chronic diseases (RIVM, 2019).

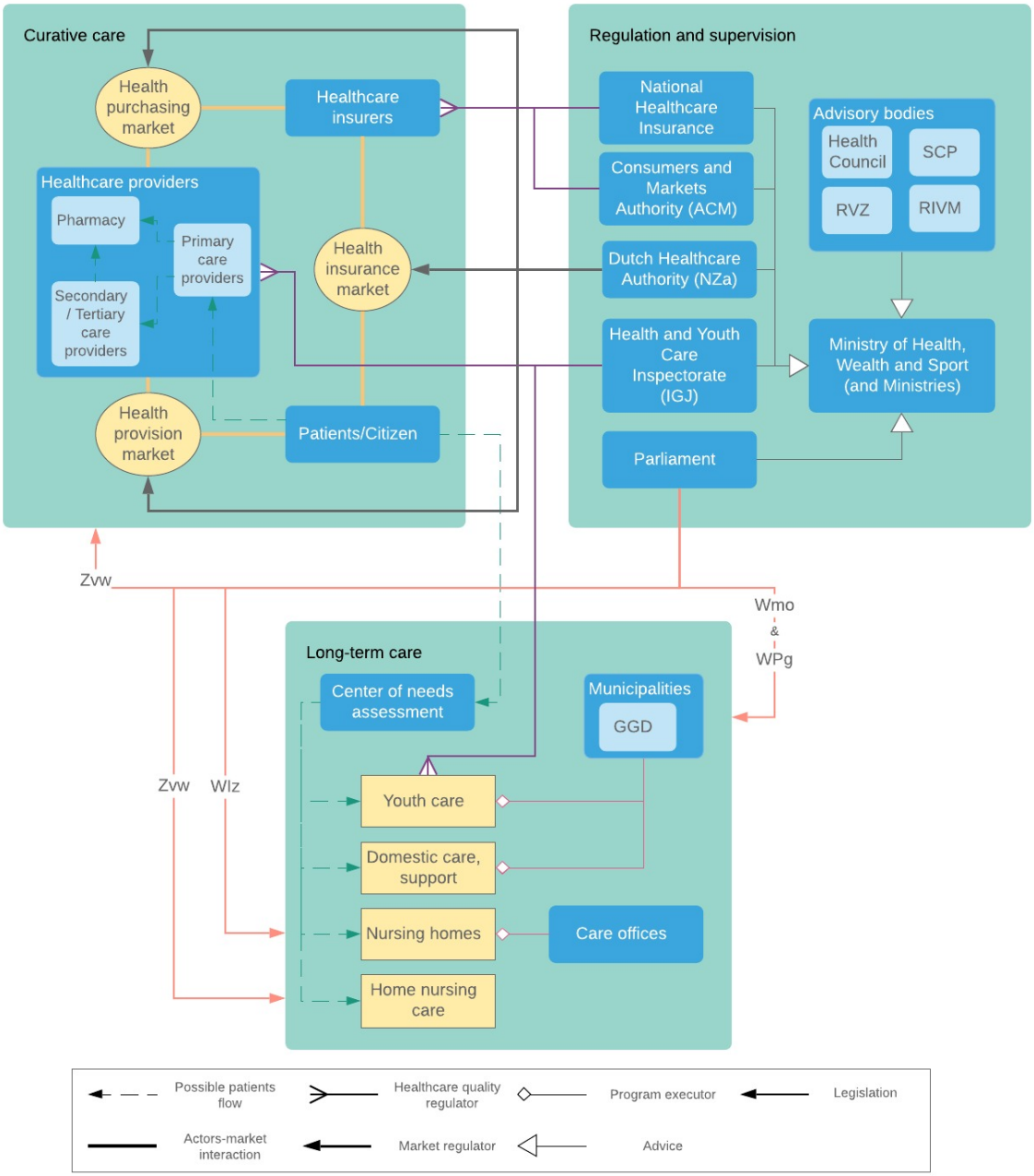


Figure 2.2: Dutch Healthcare Systems Organizational chart.

2.3. Street Level Bureaucrats in Healthcare Systems

Prevention of cardiovascular disease often referred to as a collection of activities that served the purpose of mitigating or minimizing the consequences of cardiovascular disease and disabilities from either individual or population level actions (Piepoli et al., 2016). However, the increasing risk factors among the diverse population makes it harder to do an act of prevention for cardiovascular disease(Hippisley-Cox et al., 2009). As the gateway between patients and healthcare services, GP plays a crucial role in cardiovascular prevention, especially in deciding which treatment the susceptible population should be recommended. Therefore, GP is also considered as an important decision-maker in the case of cardiovascular disease prevention. According to Lipsky (2010), public servant or someone who is capable of enforcing some policy action or interact directly with the citizens can be considered as street-level bureaucrats. He stated that there are several characteristics that shared among these street-level bureaucrats, such as ample amount of freedom in their individual interactions with clients; the need to increase the efficiency of their line of work; working under scarce resources, be it material or not; working with and in ambiguous situation or rules; and lastly the demand for their service is always less than the supply of the personnel that they have (Checkland, 2004; Weatherley and Lipsky, 1977). Hence, GP in their line of work can also be considered as one of the street-level bureaucrats. Consider the case of cardiovascular disease prevention and care, general practitioner (GP) plays an important role as they are responsible for the primary care of a patient. From the section 2.1, at least there is three policy reform (Health Act, Public Health Act, Generalist Basic mental Health Care) that mandate several medical cares to the GP, which covers not only primary care but also for several cases of mental health. On average, GP spends about 10 minutes for each consultation with a patient, which they need to make decisions about the condition of a patient (NHG, 2019a; Zha.nl, 2019). In addition to that, every GP also responsible for multiple patients with different problems, thus makes it even harder for the GP to make a judgment about a patient is given the limited knowledge, information and time that they have.

Almost all GP in the Netherlands is part of the Dutch General Practitioners Association (*Nederlands Huisartsen Genootschap*, NHG). NHG handle most of the research, information, guidelines as well as a handbook for the GP related to the care that they need to provide to the patient (NHG, 2019b). Despite the guidelines and information that NHG provides, not all GP capable of utilizing the knowledge resources available. It was reported that there are various barriers for the GP to use the guidelines set by NHG or other regulating bodies (Lugtenberg et al., 2009). For example, in the recent cardiovascular guidelines it is recommended to use systolic blood pressure instead of diastolic blood pressure for a patient with hypertension, but some GP still reluctant to use these recommendations since they are used to use diastolic blood pressure in the past. There is also a problem with interpreting the cardiovascular risk of some patients among different ethnic minority groups, that is not addressed by the guidelines. Beside unclear and ambiguous guidelines, there is also a problem related to patient behavior, such as not following the medication as suggested, not willing to take the measurement that suggested, not attending the follow-up sessions and many more. Therefore, some GP most often used their own judgment or developed their own mechanisms to do their job since the real interaction and practicality of dealings with the patients require them to do so. All of the points mentioned above sum up how GP can be considered as a street-level bureaucrat. Although GP is not directly employed by the government and work under the different healthcare providers, Lipsky argues that the term *bureaucrat* is not limited to simple formal authority employment status (Lugtenberg et al., 2009). By defining GP as street-level bureaucrats, several things can be deducted for the sake of this research. Firstly, by Lipsky definition, the framework that used or applied in the policymaking can also be applied or compared in this level or in other sectors. Secondly, the necessity of using expert judgment to determine which variables to include in the model is also important to be considered, as GP in this particular research are directly responsible in communicating the cardiovascular risk to the patient. And lastly, variables that are chosen to be used in the risk prediction model in this research will also need to consider the practicalities and efficiencies in terms of GP getting the information from the patients. For example, in using socioeconomic status, there are a lot of measures that can be used, such as income, occupation education, and

wealth. In the case of income, it is possibly impractical to ask directly to the patients on how much money they make every year or month because it is considered impolite to such a question. Instead of asking such direct question about income, GP can use some cut off income questioner, such as ”are you making more than 2000 euro per month?” would probably more appropriate rather than asking specific and detail on how much they are making. Therefore, the next section will try to explore various variables that might be worth considering to be included in the risk prediction model that is going to be constructed in this research.

2.4. Cardiovascular Risk Prediction Model

Currently, there are several cardiovascular disease risk estimation system that is commonly used in the clinical practice to provide information for cardiovascular diseases prevention such as PROCAM, SCORE, CUORE, Framingham, QRISK1 and QRISK2, ASSIGN – SCORE, Pooled Cohort Studies Equations, Globorisk. All of these risk estimation systems used different predictors and predicted variables to design the recommended treatment for primary care of potential cardiovascular disease patients.

Overall, the different risk estimation system can be split up into four different layers of categories based on the predictors that they used. In the first layer, there are biological characteristics such as sex, age, race, and BMI (body mass index). The second layer consists of measurement and medical records such as smoking status, diabetes mellitus (DM), total cholesterol, high-density level cholesterol (HDL-C), systolic blood pressure (SBP), family history and chronic disease. In addition to that, medication and medical treatment such as hypertensive, antihypertensive treatment, and BP (blood pressure) treatment. Lastly, the layer consists of social factors such as area-based social deprivation and ethnicity. Although, all of the risk systems contains the combination of factors from the first layer and second layer, not all of the system include the medication and the social factors inside the calculation. Across eight different risk framework, only 4 of them include the third layer, Framingham (D’Agostino et al., 2008), QRISK1 and QRISK2,(Hippisley-Cox et al., 2011,0,0) Pooled Cohort Studies Equations (Goff et al., 2014), and CUORE (Giampaoli, 2007). On the other hand, the social factors only appear in two risk estimation framework, QRISK and ASSIGN - SCORE (Lewsey et al., 2015). Besides that, they also used at least 10-years (or lifetime risk) risk prediction, and most of them measured a different type of risk starting from first cardiovascular disease events, fatal events of cardiovascular disease, even mortality associated with the cardiovascular disease risk. They also specify a different age range that exists between 20 to 84 years old. It is also worth mentioning that each of these frameworks was constructed with a different mathematical model, called the survival model. There are two different survival model that is mostly used in the risk estimation systems, called the Weibull model and Cox-proportional hazards. Each of these models has its assumptions and heavily dependent on the dataset that they collected during the experiment (or study). Therefore, the risk estimation result could only be interpreted in a specific region or country where the studies were conducted. For example, the most used risk estimation systems, Framingham studies were conducted in USA, SCORE and ASSIGN as well as QRISK are conducted in England and Scotland. These risk estimation systems have the primary dataset that is mostly based on the Caucasian population and has to perform better for that particular population race, but the for the non-Caucasian, it is still unknown whether this scheme is accurate or not (Piepoli et al., 2016). For more detail, the explanation about the model will be explained in chapter 3.

2.4.1. Biological characteristics

It is a well-known fact that the function of survivability of a certain individual depends on age since the body deteriorated the older a person is. In every case of risk estimation systems, age has been the primary variable used in the risk model that they built (Piepoli et al., 2016). Although every model used different age range for their model, the author found no evidence of the assumptions behind the chosen age range. Therefore, it is assumed that the age range that is chosen for the model were mainly due to data availability or limited organization of the study that was conducted.

On the other hand, gender has also been considered as important factors in predicting cardiovascular events mortality and morbidity. Literature studies found a consistent gender-specific cardiovascular risk in the risk estimation systems (Piepoli et al., 2016). Apparently being male have a higher chance of having a cardiovascular disease than female. Although each of the studies has different number about the elevated risk of cardiovascular disease for gender, the difference among the gender could be estimated around 30 - 50 percent more in all of the previous studies (Conroy et al., 2003).

There is also body mass index (BMI) [weight (kg)/height (m^2)] which act as an indirect indicator of obesity or considered as a proxy for cholesterol measurements (Hippisley-Cox et al., 2009). Literature suggests that BMI can provide to be useful as an indicator of some problem with cardiac comorbidity or some circulatory problems in the body, which could lead to cardiovascular disease (Poirier et al., 2006). Although it is not as powerful as the cholesterol measurements, It could provide to be useful when the other measurements are missing, and it is also quite easily measured compared to cholesterol level (Piepoli et al., 2016).

Lastly, there is a race which is used in Pooled Cohort Studies Equations (Goff et al., 2014) which try to compare different cardiovascular risk among the population of Caucasian and African American race. Goff et al. (2014) found out that African American have a higher chance of experiencing cardiovascular events compared to Caucasian. It is worth to mention that this research tries to distinguish the definition of race and ethnicity since in some literature they used interchangeably. According to Sheldon and Parker (1992) ," race is a biological concept which categorizes humanity by means of sets of phenotypical features that appear to distinguish between varieties of people and are passed on between generations". While ethnicity is defined as "... shared cultural characteristics and national identity". Ultimately, the race and ethnicity try to explain the same thing, which is biological characteristics that passed down by their predecessors. The only difference lies in the pass down of culture among these different race. Some people might have "Asian" or "Indian" as a race, but because of the how a person adapts to the culture and their national identity, that person could be considered as a "Dutch".

2.4.2. Measurements and medical records

There are two measurements that are used as the main indicator for a person that potentially have cardiovascular disease, cholesterol, and blood pressure. There are two types of blood pressure measurement, systolic that defined as a measurement of blood pressure when the heart muscle is pumping oxygen-rich blood into the blood vessels; while diastolic refers to the measurement of the relaxed heart muscle and when blood is refilled with oxygen (Kleinert et al., 1984). Given the time and situation when the blood pressure is measured, it can vary heavily. For example, a person can have elevated systolic blood pressure (more than 140 mmHg) whenever the person was previously exposed to some extreme cold, heat, pain, or stress. Therefore, it is important that health practitioner to measure the blood pressure multiple times or at least makes a note about the condition when the blood pressure is measured. Although almost all of the risk prediction model mentioned above use systolic blood pressure (Piepoli et al., 2016), diastolic blood pressure also used in the NHG guidelines as also a previous proxy of elevated blood pressure.

The cholesterol level is one of many causes of blockage in the circulatory body that is the main causes of cardiovascular disease (Birtcher and Ballantyne, 2004). Please note that cholesterol and fat are not the same in the medical terms even though it has a similar chemical compound. Cholesterol is produced by the body to help build cells and a membrane that is essential for the production of different substances in the body while fat is a concentrated source of energy that play an important part in chemical and metabolic functions. There are four types of variables that usually used in the risk prediction model related to cholesterol level, namely low-density lipoproteins (LDL), high-density lipoproteins (HDL), total cholesterol, and triglyceride. The amount of cholesterol that exist inside the blood (lipoproteins) is what usually measured by health practitioners. The excess amount of LDL in the circulatory systems will build up the amount of plaque in the blood vessels. Then the body develops HDL to counter the excess growth of LDL, which is a substance that able to dissolve LDL. Since

most of the food contains a higher level of LDL compared to HDL (especially greasy food), therefore it is understandable that people that higher level of LDL and/or lower level of HDL has a higher chance of cardiovascular disease. On the other hand, total cholesterol measure all of the measured cholesterol level, starting from high-density lipoproteins, low-density lipoproteins, and as well as very-low-density lipoproteins. Lastly, triglyceride measures the amount of regular fat that is stored inside the blood. It is worth to mention that each of these measurements requires the person to be fasting for the last 9 - 12 hours since the amount of LDL, VLDL, and triglyceride measurements will drastically influence by this. In all of the risk estimation system, HDL and total cholesterol have been the standard variable or indicator for a susceptible individual with cardiovascular disease. There is only two risk model that used LDL as their variables, PROCAM (Assmann et al., 2002) and QRISK2 (Hippisley-Cox et al., 2011,0) while there is no mention of triglyceride in the standard risk estimation model.

Lastly, there is a comorbidity (Valderas et al., 2009) of an individual that could lead to cardiovascular events. Variable such as Diabetes Mellitus (DM), family history of CVD, or other histories of chronic disease can be considered for the model. Despite several mentions of DM in the literature (Piepoli et al., 2016), the author only found the small portions of the risk prediction model and literature that used family history as a variable inside the model (Hippisley-Cox et al., 2011,0). Besides that, smoking status or smoking history can also be an early indicator of the susceptible individual for cardiovascular diseases. Studies suggest that smoking increase the risk of CVD by two folds, acute thrombosis of narrowed blood vessels as well as the degree of growth of atherosclerosis in circulatory system (for Disease Control et al., 2010). Therefore, in all of the risk estimation systems, smoking status always appears and considered in the CVD risk model(Piepoli et al., 2016).

2.4.3. Medication an treatment

In cardiovascular prevention, there are two types of treatment that usually used to treat a susceptible patient with cardiovascular diseases, namely lipid-lowering drugs and (anti) hypertensive drugs (Pouwels et al., 2016). Most of the lipid-lowering drugs are prescribed by health practitioners based on the measurement of the cholesterol level of a patient. These type of treatment has proven to be effective in preventing coronary heart disease events (Pignone et al., 2000). On the other hand, (anti)hypertensive treatment are prescribed by looking at the blood pressure measurement of a patient. Both of these types of medications has been widely used not only as an indicator of a change in patient measurement (blood and cholesterol) but also used as an early indication of a previous cardiovascular disease event. For example, Bandyopadhyay et al. (2015) used the medication as dependent variables for the measurement of systolic blood pressure and LDL measurement in their model, whilePouwels et al. (2016) used medication to exclude patients with previous cardiovascular events since it is not supposed to be included as part of their research. Therefore it is useful to get the information of a patient medication history since it can be linked to their cardiovascular risk. And it is also worth mentioning that medication has also been used in the existing risk prediction model that available in the guidelines, such as Framingham (D'Agostino et al., 2008), QRISK1 (Hippisley-Cox et al., 2007) and QRISK2 (Hippisley-Cox et al., 2011,0), and Pooled Cohort Studies Equations (Goff et al., 2014).

2.4.4. Socio-economic factors

Considering all of the variables that most of the scheme used, this literature research identifies the typical pattern of what current knowledge gap related to risk estimation of cardiovascular disease, that is the increasing demand for integrating of socioeconomic variables into the cardiovascular risk estimation system Hippisley-Cox et al. (2011). The earlier description indicates that multiple variables that are used across different risk estimation systems vary among different studies. Literature suggests that adding more layer of factors into the risk estimation improve the performance of the risk prediction rather than stick with only first and second layer (Lewsey et al., 2015). As can be seen in ASSIGN - SCORE and QRISK scheme, socioeconomic factors, such as ethnicity (Bos et al., 2004; Perini et al., 2018), and social deprivation (Hippisley-Cox et al., 2011) have an impact on the risk of cardiovascular disease. In ASSIGN - SCORE, for example, it is reported that adding "area-based social

deprivation" index into their factors for calculating CVD risk have outperformed Framingham risk estimation systems (Lewsey et al., 2015), same applies to QRISK (Hippisley-Cox et al., 2011) which also add ethnicity on top of it. The "area-based social deprivation" index that used in this case related to Peter Townsend (Townsend, 1987) definition of deprivation that comes from two forms, social and material. Social deprivation relates to the interaction and relationships of an individual with workplace, neighborhood, community, and family. While the material deprivation is related to the accessibility of an individual towards certain facilities, goods, and services, the measurement of social deprivation can be done in a combination of multiple variables, starting from unemployment, overcrowded housing, homeownership, household composition, income, education level, marital status, and residential mobility (Pampalon et al., 2000). All of these variables rely heavily on the context (where and when the studies conducted) therefore need to be studied carefully while the values should be also be normalized and standardized to converge into a meaningful result that can help connect the factors into public health and welfare issues. But in the end, the literature suggests that socioeconomic factors can be associated more with cardiovascular disease through by behavior-related risk factors such as dietary habits, smoking, and exercise compared to the psychological such as blood pressure or cholesterol level (Squires, 2000).

Ethnicity as a measure of Socioeconomic status

Following the definition of ethnicity that discussed previously by Sheldon and Parker (1992), this variable plays a role in determining a person eating habits. Miner et al. (2014) discussed thoroughly different ethnic minority groups by its common diet, habits, and medical aspects. For example, the habit of alcohol and smoking consumption was found to be quite high for Albanian, which highly increases the risk of cardiovascular events. There is also an indication of some ethnic minority to low income, thus forbidding them to get some health care access, also restricting them to do some physical activities that are necessary for their body (Miner et al., 2014).

income as a measure of Socioeconomic status

The literature argues that even though income, education, and occupation correlated for a measure of socioeconomic status, they do not completely overlap (Potvin et al., 2000). Income can determine individual access to healthcare services as well as dietary habits, which is relevant for the occurrence of cardiovascular disease. There are two types of income that usually reported in the study, individual income as well as households income. In most cases, adjusted households income by family size is used since there is an individual that has almost non-existent income such as housewife as well as student, kid, etc (Galobardes et al., 2006b). However, there is also a type of income that used to determine the actual income that they earned, that follows into two categories, disposable income, and gross income. Ideally, the disposable adjusted households income is used in a lot of epidemiological studies (Galobardes et al., 2006a). Income can also be measured or reported as levels of poverty since it also shows the relationships within the healthcare access of an individual. Poverty is important to consider apparently there seem to be direct relationships with low socioeconomic status with people that live below poverty line (Squires, 2000).

Occupation as a measure of Socioeconomic status

On the other hand, the occupation could also be used as a measure of indicators for socioeconomic status. It reflects individual's place within society, the stress associated with the work, autonomy in choosing health benefits, and in some cases related to exposure to some toxic environment (smoking or even nuclear radiation for example). Occupation is strongly related to income, social standing, and intellect. One of the main limitations of the occupation is that a person that is not working during the data is taken could be underestimated in the analysis (Squires, 2000).

Education as a measure of Socioeconomic status

Education reflects the early life and long term influence in-person health. For example, if an individual constantly ill in early life, this could limit person attainable education. However, education could also affect the cognitive ability of a person to communicate with the

health practitioners, therefore, giving more clear information about their health problems (Galobardes et al., 2006a). On the other hand, It is argued that person that have high education is more receptive towards the CVD risk factors (Squires, 2000). Awareness of CVD risk factors is important in CVD prevention as it is correlated with their behavioral risk factors.

The measurement of education can be done by either a categorical variable or continuous variable. The categorical usually reported as the highest degree that individual posses or milestones of specific education such as yes or no for completing university. However, a continuous variable can be reported as time spent on completing the education usually used as a measure. The drawbacks, when using education as an indicator for socioeconomic status, shows when there are different age cohorts across such a long period of time. The bias comes from when there is old generation since it will appear as though they are less educated than the younger generation even though they have different social norm given the time they are living (Squires, 2000).

Assets and wealth as a measure of Socioeconomic status

Arguably, wealth as a combination of income and total assets considered to be a better indicator of a person socioeconomic status related to their health. This is due to there is much more information contains in wealth, such as the family history of socioeconomic status, income and in some cases it can also tell the social health plan that related to pension and insurance. The wealth importance change over the life course since people with older age can have no income but with an enormous amount of wealth accumulated either financially or physically (house, cars, inheritance, investments, pension, etc.) (Squires, 2000).

Household condition as a measure of Socioeconomic status

Lastly, there are several things that can be deducted from the condition of the household, either physically or non-physical. Physically is related to materials that some particular household relate to, such as the house condition (households amenities) such as air pollution around the living place, clean water access, heating system, housing tenure and many more. While non-physical relate to household composition, marital status, position inside a household, a number of siblings, house tenure (house ownership - rent, own with a mortgage, family own, social housing, etc.). The physical household condition can be linked to a specific condition of health outcomes while some non-physical can be subtlety link to socioeconomic status to predict some occurrence of some complication of disease (Squires, 2000).

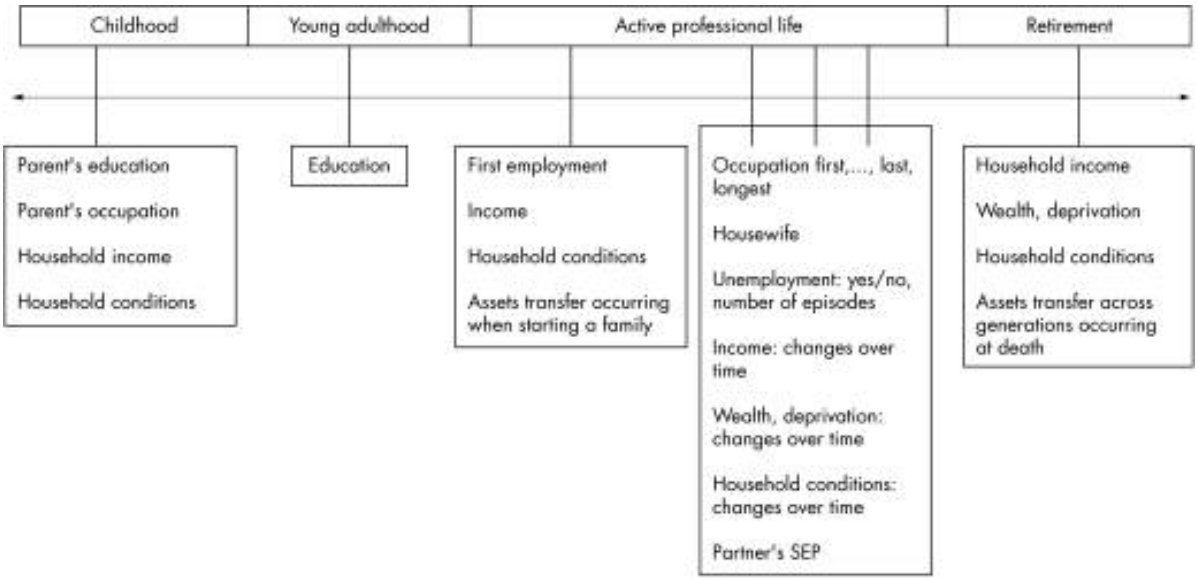


Figure 2.3: Socioeconomic status timeline (Source: Squires (2000)).

Concluding remarks about this subsection, it seems that every socioeconomic factor can have different information and story when included in predicting some certain health out-

Figure 2.4: Summary of all variables mentioned in the literature.

come (summarized by Figure 2.3). Literature mentioned the connection between socioeconomic status to alcohol and smoking consumption's; dietary habits, thus BMI (Méjean et al., 2013; Psaltopoulou et al., 2017); and connection between ethnicity and the income and wealth (Psaltopoulou et al., 2017; Squires, 2000). Figure 2.4 Summarizing all of the variables mentioned in the literature for cardiovascular disease risk.

3

Methodology

There is various modeling method to calculate the cardiovascular risk of a person. In the context of this research, the two most common survival analysis will be explored, the statistical method and machine learning method. There are three types of statistical method that popular in the survival analysis and will be used in this research, non-parametric, parametric model and semiparametric model. While the machine learning method that is going to be considered in this research going to be Bayesian network. Wang et al. (2019) gives a summary of multiple survival analysis that is most commonly used and also the recent developments of the survival analysis technique. This chapter will explain several basic functions of the model as well as several benefits and drawbacks of each modeling approach.

3.1. Censoring and Truncation

Before proceeding into the definition of each different model available, there are several definitions that need to be explained related to the several assumptions about the information that is contained in the dataset. The most common term used in the survival analysis is called ”censoring” and the second term is ”truncation”. First, the right censoring happens when an individual has not experienced an event of interest after the study ends. On the other hand, when an individual is considered to have an event of interest before the study start, this belongs to the definition of left-censored. Interval censored occurs when an individual is lost to follow up during the study and occurs again before the study ends, and therefore, the information about the survival time during those missing period is unknown. Besides that, there are left and right truncation, which is to study specific issues that try to exclude some of the subjects that have specific characteristics before or after the study. Truncation usually occurs related to a study that analyzes some specific age groups during the period. So for example, If the study chooses to select the only population that has aged more than 35 years old, then the data is left truncated on age lower than 35 years old and vice versa for the right truncated data. Most of the statistical method for survival analysis can take into account the right-censored data while the left-censored or left truncation is often unknown and most often neglected (Cain et al., 2011). Therefore, for the following section, when censoring is discussed in the method, it is mainly referring to right-censored data and not other types of censored data.

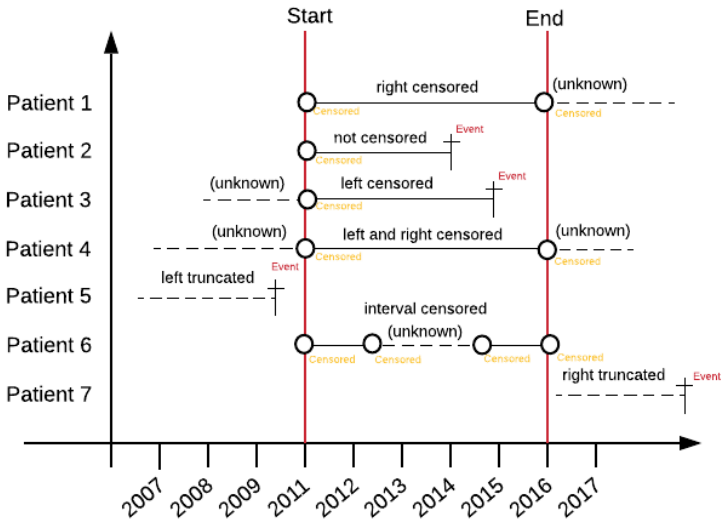


Figure 3.1: Censoring and truncation illustration

3.2. Metrics for survival analysis

There are two definitions that used to measure the time-to-event risks of an individual, survival rate, and hazards rate. The survival rate is the rate of population or subject of interest survives given the period that was chosen for the study (Harrell Jr, 2015). It is defined by the survival function (Function 3.1) with T as the non-negative value that represent the lifetime of an individual before time of interest which defined as t. Thus, the survival function $S(t)$ is defined as the probability of an individual to survive before time t with $F(t)$ is defined as cumulative distribution function of T.

$$S(t) = \text{Probability}\{T > t\} = 1 - F(t) \tag{3.1}$$

On the other hand, hazards rate is defined as the likelihood of subject experiencing the instantaneous events overtime (Harrell Jr, 2015). The hazards rate can be formally represented as the hazards function $\lambda(t)$ that is derive from Function 3.1 as the following:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \tag{3.2}$$

or in much simpler terms define as,

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt} [\ln S(t)] \tag{3.3}$$

with cumulative hazards function define as,

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{3.4}$$

or in terms of survival function,

$$\Lambda(t) = -\log S(t) \tag{3.5}$$

$$S(t) = \exp(-\Lambda(t)) \tag{3.6}$$

3.3. Statistical method

Only several models that are most commonly used are going to be considered for each type of survival analysis with the statistical method. First, in non-parametric, Kaplan-Meier model is used while for semi-parametric this research used the Cox proportional hazards model. Then, Exponential, Weibull, Gamma, Log-logistic and Gompertz are chosen for the parametric. The Figure 3.2 visualize the assumptions of each model type while the potential for different type of statistical method are summarize in Table 3.1.

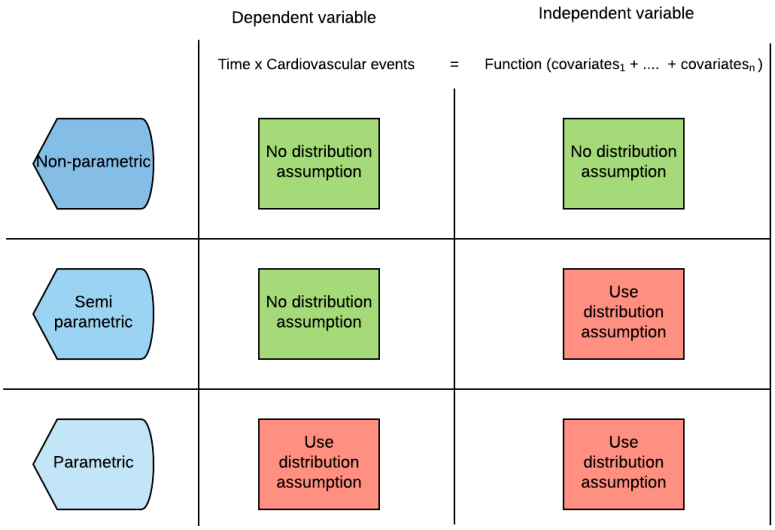


Figure 3.2: Overview of statistical survival analysis

Table 3.1: Opportunity of Statistical Method

Statistical method	Advantages	Disadvantages
Non-parametric	More efficient when nosuitable theoretical distributions known	Difficult to interpret;yields inaccurate estimates; not possible to add covariates
Semi parametric	The knowledge of theunderlying distribution of survival times are notrequired	The distribution of the outcome is unknown; noteasy to interpret
Parametric	Easy to interpret, moreefficient and accurate when the survival timesfollow a particular distribution	When the distribution the assumption is violated, it may be inconsistentand can give sub-optimal results

Source : Wang et al. (2019)

3.3.1. Non-parametric model

In non-parametric survival analysis, the survival curve is not assumed to follow specific probability distribution hence make no assumption about the shape of the survivability of the individuals in the study. The most common application of the non-parametric model is the Kaplan-Meier method capable of estimating survival times that can make a proper estimation for censored individual Goel et al. (2010). The survival function of of Kaplan Meier estimate are described in Function 3.7 with d_i as number of individuals with an event at time t and n_i as the number of individuals at risk at time t. Given that there is a censored individual (c_{j-1}), then the number of people at risk (n_i) will be defined as $n_i = n_{i-1} - d_{i-1} - c_{i-1}$.

$$S(t)_{\mathrm{KM}} = \prod_{i:t_i \leq t} (1 - d_i/n_i) \tag{3.7}$$

Based on Function 3.7, there is no notation about covariates in the formula. Therefore, usually when trying to compare the effect of covariate into the survival rate using the Kaplan-Meier model, each unique values in the covariate need to be modeled separately. However, as

the number covariates increases, the difficulties of interpreting the result become apparent, especially if the covariate is continuous variable (Clark et al., 2003).

3.3.2. Semi parametric model

Cox Proportional Hazards is considered as a semi-parametric model since it can accommodate analysis without underlying assumptions about particular survival time distribution while also offering the possibilities to add the effect of covariates given its distribution which non-parametric lacking (Harrell Jr, 2015). The hazards function of the Cox Cox Proportional Hazards (Function 3.8) is represented as $\lambda(t|X_n)$ which translate to the hazards of time t given the set of n covariates X. $\lambda_0(t)$ in the formula is the baseline hazards or the value of the hazard given that all covariates value are zero while β and X are the regression coefficients and the covariates respectively.

$$\lambda(t|X_n) = \lambda_0(t)\exp(\beta X_n) \tag{3.8}$$

or can be rewritten as,

$$\lambda(t|X_n) = \lambda_0(t) \times \exp\left(b_1x_1 + b_2x_2 + \ldots + b_nx_n\right) \tag{3.9}$$

However, as can be seen in the Function 3.9, the baseline hazards value are not assumed to be following a certain distribution; therefore it is important that the baseline hazards value are following proportional hazards assumptions. Formally, to be able to use Cox proportional hazards, the effect of a change in a covariate on the hazard rate of event occurrence should be constant over time (Park and Hendry, 2015). This proportional hazards assumptions usually represented by hazards ratio (HR) that is the ratio between two groups or subjects should not vary across time as represented by Function 3.10 (i and j are two different groups or subjects).

$$HR_{ij} = \frac{\lambda_i(t|X_n)}{\lambda_j(t|X_n)} = \frac{\lambda_{0i}(t) \times \exp\left(b_1x_{i1} + b_2x_{i2} + \ldots + b_nx_{in}\right)}{\lambda_{0j}(t) \times \exp\left(b_1x_{j1} + b_2x_{j2} + \ldots + b_nx_{jn}\right)} \tag{3.10}$$

3.3.3. Parametric model

The parametric model assumes that there is an underlying assumption or pattern that survival times follow. In terms of survival analysis, it tries to fit the survival times into some certain probability distribution function. It is possible for this type of model to also include the effect of various covariates into the survival times of the data. As opposed to the semi-parametric model, the parametric model is relatively simple and effective in estimating the time-to-event given that it is intended for multivariate analysis (with covariates) (Wang et al., 2019). However, the effectiveness and the correctness of the parametric distribution depends on how well the model finds the parameters of the distribution that fit into survival times. There are several well-known models that are widely used in survival analysis, as can be seen in Table 3.2. Given the number of parameters of the distribution, the flexibility of parametric distribution in predicting the survival times of the dataset can be known. If the researcher ought to perceive that the survival curve to be steep and not constant, they might consider using more optimized distribution such as Gamma or Gompertz while avoiding the use of much simpler distribution such as exponential (Bradburn et al., 2003b).

Table 3.2: Different parametric models

	Survival function	Hazard function
Exponential	$S(t) = e^{-\lambda t}$	$\lambda(t) = \lambda, \quad \lambda > 0$
Weibull	$S(t) = e^{-\lambda t^{\gamma}}$	$\lambda(t) = \lambda\gamma t^{\gamma-1}$
Gamma	$S(t) = 1 - I_k(\lambda t)$	$\lambda(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(1-I_k(\lambda t))\Gamma(k)}$
Log-logistic	$S(t) = \left(1 + at^b\right)^{-1}$	$\lambda(t) = \frac{abt^{b-1}}{1+at^b}$
Gompertz	$S(t) = e^{-\frac{a}{b}\left(e^{bt}-1\right)}$	$\lambda(t) = ae^{bt}$

3.4. Discrete Bayesian Network

Besides the statistical method, there is also the possibility to use a Bayesian Network model that is quite powerful in terms of interpretability of the model as well as the inference of the result. However, despite several success of the application of Bayesian Network in medical prognostics, the availability of the applications of BNs to medical problems still relatively few compared to the statistical method which is more popular among expert (Twardy et al., 2006). The Bayesian Network operated through the directed acrylic graph (DAG) to define the model structure and try to explain the dependencies behind the variables that are used. Each variable or also known as nodes are connected through *arc* (usually represented as an arrow for directed or line for undirected) or *edge* that represent the direct probabilistic dependencies between the node. If there is no arc connecting two nodes, then it is assumed that the variable is either independent or conditionally independent to each other. Bayesian Network also offers the possibility of including the knowledge expert about the model structure as well as putting a biased estimate in the dataset. There are three types of Bayesian network analysis that is usually used in practice, discrete Bayesian network, continuous Bayesian network, and hybrid (a combination of both). However, due to some limitation of this research, only Discrete Bayesian Network is going to be considered in this thesis.

Formally, Bayesian network are defined as a graphical models that represent a multivariate probability distribution of a set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ which denotes as Function 3.11.

$$\Pr(\mathbf{X}) = \prod_{i=1}^{p} \Pr\left(X_i | \Pi_{X_i}\right) \tag{3.11}$$

or can be re-written as the chain rule function,

$$\Pr(\mathbf{X}) = \prod_{i=1}^{p} \Pr\left(X_i | X_{i+1}, \dots, X_p\right) \tag{3.12}$$

Structure Learning

The first step into Bayesian Network modeling is to define the model structure. Which means that the dependencies between the set of random variables $\mathbf{X}$ (DAG) need to be defined. However, the difficulty of learning the DAG of Bayesian Network increase exponentially as the number of variables grows (Scutari and Denis, 2014). Therefore, usually, only a small portion of the node is usually inspected given there is a limited resource of the research. There are three methods that employed when doing structure learning for Bayesian Network, expert opinion, literature review, and using some *ad-hoc* algorithm to explore the plausible model structure. For example, Bandyopadhyay et al. (2015) used clinical knowledge of an expert to define model structure while Twardy et al. (2004,0,0) used literature to justify their model structure. In addition, there is also the possibility to compare a different plausible model that is generated through algorithms that used statistical criteria. The most common statistical criteria called the network score that focuses on measuring the statistical fitness of the generated model to mirrors the dependence structure of the data. Algorithms such as hill-climbing are one of the examples of network score based on structure learning. It begins by generating DAG with no arcs while gradually add, removes, and reverse the arc direction step by step while selecting the highest change in the network score. The most popular scores that implemented in the hill-climbing algorithm is *Bayesian Information Criterion* (BIC) which calculated by Function 3.13 with d being the number of parameters that the $\mathbf{X}_i$ holds and n being the sample size.

$$\text{BIC} = \log \widehat{\Pr}(\mathbf{X}) - \frac{d}{2} \log n \tag{3.13}$$

Parameter Learning

After the Bayesian Network model structure is known, the joint probability distribution (global distribution) of the model need to be specified. However, even for a simple model,

assigning joint probability distribution for the model will be difficult as there are multiple combinations of parameters that need to be defined. Therefore, the global distribution is distributed into smaller local distribution according to the arcs that set during the structure learning. Each node that is connecting to other arcs is assumed to have dependant probability distribution while nodes that are not connected will be assumed to be conditionally independent. Each node only depends on its parents (arc pointing downwards), and its child (arc pointed towards) is dependent only on its parent. There are two methods of calculating the local distribution of the Bayesian Network model, Maximum likelihood estimation, and Bayesian approach.

Maximum likelihood estimation is quite straightforward; it is calculating the empirical frequencies of for each unique parameters that are connected through their arcs. For example, consider a case where variable SEX is the parent of CVD (cardiovascular events). Then the local distribution of being "*Male*" and have cardiovascular event (*Event*) that is estimated as Function 3.14 which yields classic frequentist approach.

$$\widehat{\Pr}(CVD = \text{Event} \,|\text{SEX} = \text{Male}) = \frac{\widehat{\Pr}(CVD = \text{Event}, \text{SEX} = \text{Male})}{\widehat{\Pr}(\text{SEX} = \text{Male})} \tag{3.14}$$

Bayesian approach also can also be used as an alternative to calculating local conditional probabilities by utilizing the posterior distribution. The allocation of a uniform prior for every conditional probability table to the estimated posterior probability makes Bayesian more robust in terms of predictive power compared to maximum likelihood estimates (Scutari and Denis, 2014). By assigning the weight called imaginary sample size (iss) or equivalent sample size to the prior distribution, the estimated posterior is computed as the weighted mean of the empirical frequencies. Let us consider the same case as Function 3.14 with n as number of rows in the dataset, this gives us Function 3.15 and 3.16.

$$\hat{p}_{\text{ Event, Male}} = \frac{\text{number of observations for which } CVD = \text{ Event and SEX} = \text{Male}}{n} \tag{3.15}$$

$$\hat{p}_{\text{Male}} = \frac{\text{number of observations for which SEX} = \text{Male}}{n} \tag{3.16}$$

Then, the prior estimate of the distribution will be define as Function 3.17 and 3.18.

$$\pi_{\text{ Event , Male}} = \frac{1}{\text{n}CVD \times \text{n}SEX} \tag{3.17}$$

$$\pi_{\text{Male}} = \frac{\text{n}CVD}{\text{n}CVD \times \text{n}SEX} \tag{3.18}$$

Which then the weighted mean of the empirical frequencies will become Function 3.19 and 3.20. By putting these two function into Function 3.14, hence the posterior estimate of the Bayesian approach of parameter learning will conclude.

$$\widehat{\Pr}(\text{CVD} = \text{Event}, \text{SEX} = \text{Male}) = \frac{\text{iss}}{n + \text{iss}}\pi_{\text{Event,Male}} + \frac{n}{n + \text{iss}}\hat{p}_{\text{Event,Male}} \tag{3.19}$$

$$\widehat{\Pr}(\text{SEX} = \text{Male}) = \frac{\text{iss}}{n + \text{iss}}\pi_{\text{Male}} + \frac{n}{n + \text{iss}}\hat{p}_{\text{Male}} \tag{3.20}$$

4

Research framework

This chapter explored the stepwise approach or framework of data modeling that used in this research. Following the CRISP-DM framework, each section describes the steps in which data modeling was conducted. Since the first step of CRISP-DM, which is business understanding already explained in the previous chapter, the next chapter will explain more about the actual process of data modeling that is conducted.

4.1. Data understanding

There are two main sources of the dataset that used in this research, CBS microdataset and the Hague GP dataset from ELAN (Extramural Leiden Academic Network). The CBS dataset contains various demographics, socioeconomic, death record (encoded with International Classification of Disease - ICD 10) as well as some of the medicine purchase records (encoded with Anatomical Therapeutic Chemical Classification System - ATC 4). While GP dataset contains various measurement of a patient with a code of diagnoses (International Classification of Primary Care - ICPC 1) as well as prescribe medication used by the patient (encoded with Anatomical Therapeutic Chemical Classification System - ATC 6). Two website are used in the case of understanding the dataset (CBS, 2018b) for the CBS and (LUMC, 2018; NHG, 2019a,1) for the GP dataset. However, the dataset that is collected had a discrepancy in terms of the time when the data is collected. For example, in CBS dataset, even though the demographics information such as ethnicity, date of birth, address and sex was collected starting from 1990 to 2018, the socioeconomic status only available starting from 2011 to 2017. On the other hand, the ELAN data have the dataset that starts from 2009 until 2017 for all the variable.

As there are at least more than ten different variables inside each subcategory of different socioeconomic categories, the decision has to be made regarding which variable to choose from CBS dataset. For the full portion of the data dictionary, please refer to Appendix A.1.1. Thus, to narrow the variable down into even meaningful variables to be included and test in the model, the variables that listed in the Appendix A.1.2 was consulted with the data administrator (from CBS) as well as expert opinion and author assumption from various literature that discussed in the Chapter 2 producing the Table A.1. Unfortunately, according to the data manager from CBS, the all the variables that are related to occupation and education are under-reported and drawn from the really small sample set, therefore cannot be used for the research purposes. However, income and household-related variables that available in the CBS datasets are fine to use and recommended by the data administrator.

There is also a difference in terms of missing values for each variable in the dataset. For example, the person ID from the CBS dataset as well as the person basic demographics such as sex, date of birth, and ethnicity are fully filled in the dataset. However, the missing values start from a various socioeconomic variable, since it only contains data from 2011, the data from before that is missing with around 35 percent missing if it were drawn from 2007 to 2017. But if the data for socioeconomic variables are taken from 2011 to 2017, then the

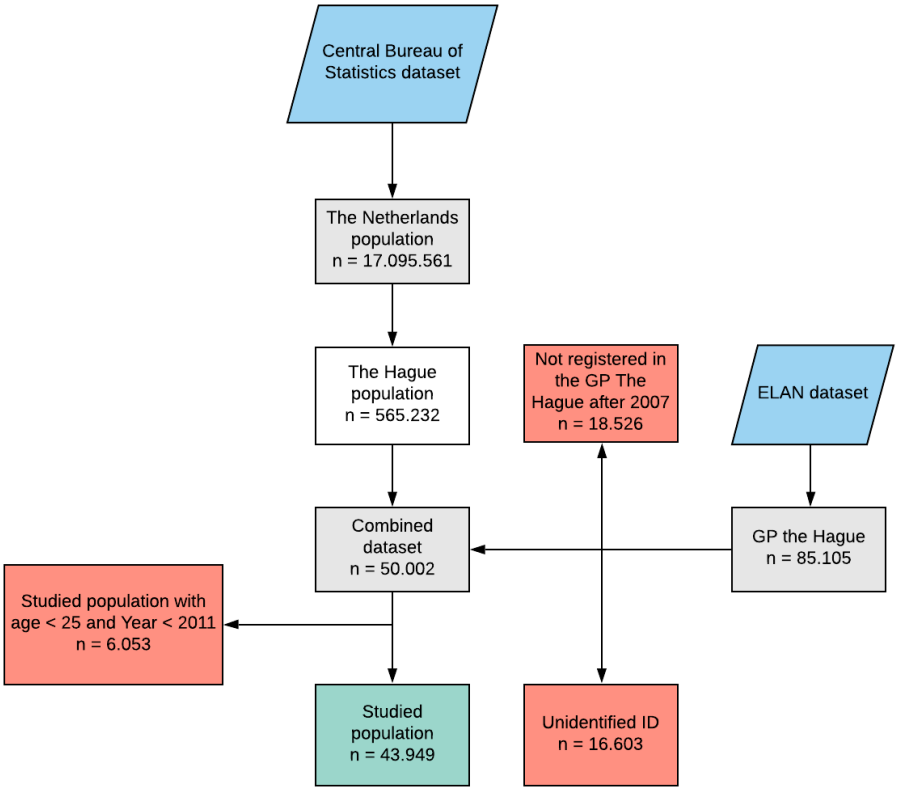


Figure 4.1: Number of individuals for each process

missing data are estimated only around 5 percent. On the other hand, ELAN dataset is even worse, with more than 75 percent missing data for all the measurement, while the diagnosis dataset only covers about 30 percent of the whole datasets. However, for the medication variable, since CBS and ELAN dataset contains medication records, it could complement each other, with CBS data cover the medication data from before 2009 while ELAN dataset covers the dataset form after 2009. The data about death, only available in the CBS dataset and fortunately available starting from 1990 to 2017. There is also discrepancy about the dataset from the date of when the data is reported. In CBS dataset, all the data are reported yearly, while for the ELAN dataset, the data are reported when there is a visit giving the more detailed date.

Lastly, the data from CBS are only available in the CBS environment (server) therefore the ELAN dataset has to pass through the data inspection from the data administrator to be able to deploy in the CBS environment. In CBS dataset, every person has their own encrypted, unique id called RINPERSOON for individual and RINPERSOONHKW. However, in the ELAN dataset, each individual has their own encrypted patient id that registered in the GP system, called gp_patidf_crypt. Since ELAN dataset was uploaded in the CBS environment, the patient general patient id was given the matching RINPERSOON that can be matched to the CBS dataset. However, it is possible that there is an individual that cannot be identified by CBS admin, therefore, reducing the dataset even less than before. For a complete overview of the number of a person that available for each dataset, please refer to the Figure 4.1 and the process of how the data is preprocessed will be discussed in the next section.

4.2. Data preprocessing

The data preprocessing consist of three phases, data cleaning and merging, missing data imputation, and data discretization. Data cleaning and merging are intended for combining the CBS dataset and the ELAN dataset. On the other hand, missing data imputation is

intended to fill in the missing values with three different data treatment, imputation with single imputation of either median or mode; change all missing values into specific value; and imputation with multiple imputation chain equation. Lastly, the discretization is performed since it is necessary for the input of a discrete Bayesian network that is going to be performed.

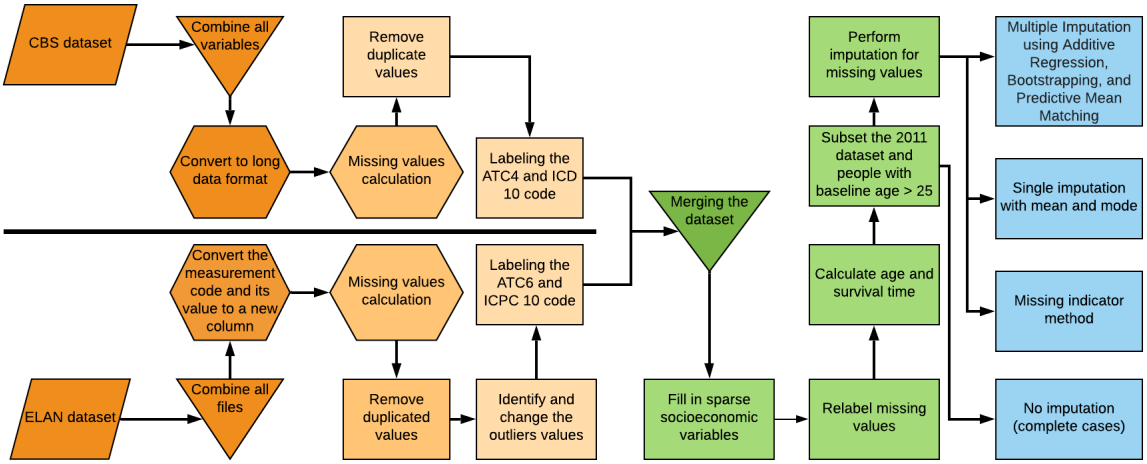


Figure 4.2: Preprocessing overview

4.2.1. Data cleaning and Merging

Before explaining the data cleaning process, it is worth to mention that the data that available in CBS has at least more than 1 GB for each yearly combined file. For example, INHATAB file (please check table A.1.1) that are combined into one file from the year 2011 to 2017 has at least 7.8 GB which consist of 24 different variables (columns) with at least 100 million rows (17 million citizen x 6 years). Additionally, the limited memory (random access memory, RAM) capacity (up to 8 GB and could be less) that is used in the CBS environment makes the matter of data handling even worse. Therefore, some early data cleaning process was done in SPSS statistical software while some more elaborate data cleaning and modeling will be done in R studio. The early data cleaning process is preferable as the data file that is used in CBS environment used SPSS standard file (.sav) and SPSS allows for the ”spillover” memory (random access memory, RAM) than capable of using the allocated hardrive memory in the case of data is too big for the memory (random access memory, RAM).

In the early stage of data cleaning, all the necessary file that contains the variable of interests need to be combined into one SPSS data file since they are located in various location in the CBS environment. Normally, data handling in SPSS uses the ”wide data structure,” which means that if there are repeated measurement or values for a particular unique row, will be represented as another column in the dataset. On the contrary, the data handling in R used a long data structure that represents the repeated measurement as a duplicate over the same unique id with the different values for each measurement (represented unique id over the year). As CBS used SPSS as the standard program that represent the data, it is also logical that the CBS data that available in the wide structure while the ELAN dataset used long data structure. Therefore, the CBS dataset needs to be transposed first (in SPSS) before proceeding to the next stage of data cleaning.

On the other hand, the ELAN dataset that already combined into one file needs a little bit data restructuring, especially for the all the measured variables in the ”MEETWAARDENCBKV1” file since the file also contains unnecessary measurement that is not needed for the measurement. Additionally each measurement in the ELAN dataset use NHG encodings NHG (2019b) which have various code for even the same type of measurement. For example, there are at least ten different systolic blood pressure measurement that available in the dataset, standard systolic blood pressure measurement (RRSYKA), systolic blood pressure (horizontal) (RRSYKAPR), systolic blood pressure (standing) (RRSYKAPS), systolic blood

pressure (target) (RRSWKQ), etc. Therefore, to select the relevant measurement code in this research, expert in the medical health (GP) are consulted and the selected code of measurement can be seen in the table A.1.2 in the ”variable names” column with the first keyword as the column name in the ELAN datafile and the keyword after as the code name. However, for the smoking status, expert suggests taking multiple codes as an indicator as opposed to a single code which is ROOKAQ for whether people smoking or not. For instance, there is a record that indicates whether a person advised to stop smoking or not, which could be an indicator that they are indeed smoking at that time. And there are other variables proxy that also into account such as a number of cigarettes a person smoke, years of smoking, consultation of quit smoking, etc. Subsequently, after knowing which code to take, the dummy column is created with the column name the same as measurement code, and all the measurement values are copied into the new dummy column.

Thereafter, outliers in the measurement dataset are explored by consulting an expert as well as NHG guidelines for determining different cut off points and logical values of measurement variables. For instance, according to NHG guidelines and expert, systolic blood pressure should normally be around 50 and 250 mmHg, and therefore, any values beyond that will be considered as outliers. Then, the outliers values that are successfully identified are going to be considered either as a missing (NaN) value or change into some logical value. Assumptions are made for changing some of the outliers values of the measurement. The first assumption is that the general practitioner put more zero by mistakes, while the second assumption deals with misreported first value by 1. For example, there are several cases of reported systolic blood pressure that is more than 1000, such as 1600 mmHg. Then the new value of this variable will be assumed to be 160 mmHg (1600 /10). The second example, for instance, for reported cholesterol level, that gives number 7.2, will be reduced to be 6.2 since it is assumed that there is a typo in inputting the value. Fortunately, in CBS microdataset there is no extreme number that goes beyond the specification of the what listed in the data dictionary; therefore, it is assumed that there are no outliers in CBS dataset.

Thereafter, both datasets are ready to be combined by matching RINPERSOON or a person unique ID that available in both dataset. Although not every data can be matched with each other and therefore, only an individual that can be matched is considered in this research. Moreover, before merging the two datasets, some of duplicates values need to be removed in the two datasets to ease the computing process as from this point on the data handling will use R studio instead of SPSS. The removal of the duplicated values are based on the two conditions, the least number of missing values and deletion of multiple duplicated values in multiple columns. First, the dummy column is created that calculate the number of missing values in each row. Then, each unique person id is ordered based on several columns of interest, such as the CVD event, medication, measurement, etc., which also include the count number of missing values. Then only the first out of multiple duplicated values are chosen as the values by using multiple columns of interest as a condition. For example, if person ID 1 is duplicated three times, then the first thing that the code does is to calculate how many missing values in each duplicated values. Then these three duplicated people are ordered based on multiple column values, for example, column ”Year, Label of death, Medication, Income, and Missing value count” therefore the first duplicated values contain the most information out of all other duplicated values. After that, the code will choose only the first out of three duplicated values. As to make sure to get the most value with the highest number of information, the removal of the other duplicated values is also double-checked with putting a condition to delete only if they have matching or duplicated values over multiple columns that used when the code ordered the duplicated person ID. Thus, in both dataset, the one value that is chosen in multiple occurring duplicates are considered to represent the most information possible out of other duplicated values. After the duplicated values are removed, the dataset can be merge by using person ID or RINPERSOON as the key to match.

After the ELAN and CBS dataset are merge into one, the difference time when the variable values are reported from both dataset makes the data really sparse. Therefore, for the socioeconomic status, it is assumed that given there is repeated measurements value every year, the value of socioeconomic variables will be the same for that particular year. For

example, if there are at least three recorded medical measurements for patient 1 and only one recorded income values for the year 2012, the value of other two recorded medical measurements for the income will be assumed to be the same. However, this assumption only applies to the socioeconomic status, and the medical measurement values are not following these assumptions since it contains more dynamic state compared to socioeconomic status variables. Thereafter, some simple statistical summary and frequency count of the dataset is performed to check the quality and the other type of missing values in the dataset. The summary of the data can be seen in table 4.1. As can be seen, the dataset contains a lot of missing values given the time period that is chosen. Given a piece of information that the value holds, there are two types of missing values that identified upon closer inspection about the dataset, known missing values and unknown missing values. As the name suggest known missing values are values that labeled explicitly as unknown or missing, while known missing values are missing values that do not carry specific label values. For example, for variable INHBESTINKH, which is disposable (spendable) household income of an individual, CBS put values of ”999999999” for the household without known spendable income. On the other hand, measurement dataset also has similar issues with known missing values, in variable smoking status, which labeled as either ”unknown” or ”not known”. However, both types of missing values are fundamentally not informative; therefore, the known and unknown missing values need to be re-coded or relabelled as the same ”NaN” values.

Then, it is decided that for this research to use the data from 2011 to 2017 given that the 2011 dataset contain the least percentage of missing values compared with other datasets. Therefore, The data is *left censored* from 2011 backwards and *right censored* forward from 2017. Thereafter, the calculation of age and follow up time are performed. There are two types of age that are calculated, the dynamic age and baseline age. The dynamic age refers to the difference between the date of birth and calendar time of the recorded measurement. While the baseline age refers to the difference between the date of birth and the calendar time of the start of the study, which in this case, refer to the 1st of January 2011, on the other hand, the follow-up time refers to the time elapsed from the start of the study until the end of the study (31th of December 2017) with year as a unit. Afterward, the dataset is subsetted for only data that exist after 2011 and individual that have aged at the baseline more than 25 years old. The chosen age cut off is deemed to sufficient as it is highly unlikely that the data will contains some cardiovascular events at a younger age. And even if it exists, it is likely to be caused by genetic or only in some rare group of the population that is outside the scope of preventive measures of cardiovascular disease.

Table 4.1: Data summary

Information	2007	2009	2011
Number of observations	611,114	494,195	377,885
Number of Person	46,296	45,722	43,949
All Death	2,333	2,237	1,712
CVD Death	469	445	348
First CVD Diagnosis	4,386	3,905	2,652
CVD Events (ICPC + ICD)	4,847	4,343	2,994
Missing values CBS dataset (%)	34 - 36	22 - 23	4 - 6
Missing values ELAN dataset (%)	80 - 89	71 - 85	67 - 82

4.2.2. Missing data imputation

Even with all of the data cleaning process, there are still some missing values which could potentially reduce the number of data points that are going to be used in data modeling. Although it is common in the epidemiology studies to have missing values in the dataset, literature (Donders et al., 2006; Pedersen et al., 2017; Sterne et al., 2009) suggests a various method for dealings with missing values given the assumptions about the missing values. The three most often assumptions related to the type of missing values that most often categorized as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The first term refers (MCAR) to missing values that exist without any log-

ical explanation or unforeseen circumstances between the missing values and the observed values. For instances, missing values that exist because there is a period of time when the blood pressure measurement or cholesterol devices is not working or even destroyed during the general check-up period. On the other hand, MAR defines the existence of some of the missing values can be explained by the observed values. For example, the data of some blood pressure measurement will likely to be missing among the younger group of the population simply because they do not see the necessity to have to checkup as they think they are healthy. Lastly, MNAR explained the likelihood of missing values occur because there is a dependency between the observed values and unobserved variables. For instances, people will less likely miss the measurement appointment if their sickness feeling is getting worse.

Given those assumptions of missing values, several methods are suggested by the literature, namely, complete cases analysis, missing indicator method, single value imputation, and multiple imputations. Firstly, complete cases analysis used only the value that has all the observed data, which then exclude or omit all the recorded data that contains missing values. Secondly, the missing indicator method tries to change and re-code all the missing values into some specific unique value, and in this research, 0 or ”missing” will be used. Then, there is a single value imputation which changes the missing values by taking the most occurring values in the dataset, either mean or mode for categorical variables. Lastly, there is a multiple imputation method that used an ensemble of algorithm and sampled based on the distribution of multiple imputed datasets. The purpose of the multiple imputation method is to ”predict” or ”match” the likelihood values of the missing values given the other observed variables. As imputation is not one of the scopes of this research analysis, the detail explanation of it is omitted from this thesis and for more detailed step by step of multiple imputations, please visit Van Buuren (2018); White et al. (2011) and Rubin (1996); Siddique and Belin (2008). Each of these models has its own assumptions regarding the type of missing values that exist in the dataset. Pedersen et al. (2017) mentioned the benefits and drawbacks of implementing the different method, and this research acknowledges these limitations of every method. Therefore this research considered using all the method mentioned above and used each different data treatment for missing values as a consolidation of data modeling purpose for selecting socioeconomic status.

The implementation of some method of imputation for the missing values is slightly altered in this research. For instances, the single imputation with mean is performed by the overall mean or mode of the whole datasets but instead performed based on every subset of repeated individual measurement. In addition to that, if there are completely missing values of that particular individual, then the mean or mode are taken by the subset of the values for the subset of data of the particular year when the data is missing. For example, if an individual 1 missing an income measurement in the year 2013, the mean of income of individual one from 2011 to 2017 are calculated (excluding the missing), and then the result will replace the missing values. However, if the data is completely missing for individual one income, then the mean of income of each year from the overall dataset is used for filling the missing values. On the other hand, multiple imputations are performed using R package ”Hmisc” with ”aregimpute”. Then, three auxiliary variables to predict missing values are used, age, sex, and ethnicity. In addition to that, all variables that contain missing values are also used to act as an intermediary to predict the missing values. Then, 5 number of imputation and with 0 number of knots (assumed to be linear) decided to be used since it is recommended in the literature (Pedersen et al., 2017; Van Buuren, 2018; White et al., 2011). As can be seen in Figure 4.3, the shape of the data with and without the imputation is not that much different. Moreover, the data treatment for the missing indicator is done in a straight forward manner by changing the ”NaN” values into 0. However, the complete case analysis can only be perform if the covariates only include the socioeconomic status since the complete data is 94 percent (approx. 360,000 observations) while if the analysis includes measurement, then the analysis will only cover 18 percent (only approx. 28,000 observation) of the dataset which is pretty low for the analysis.

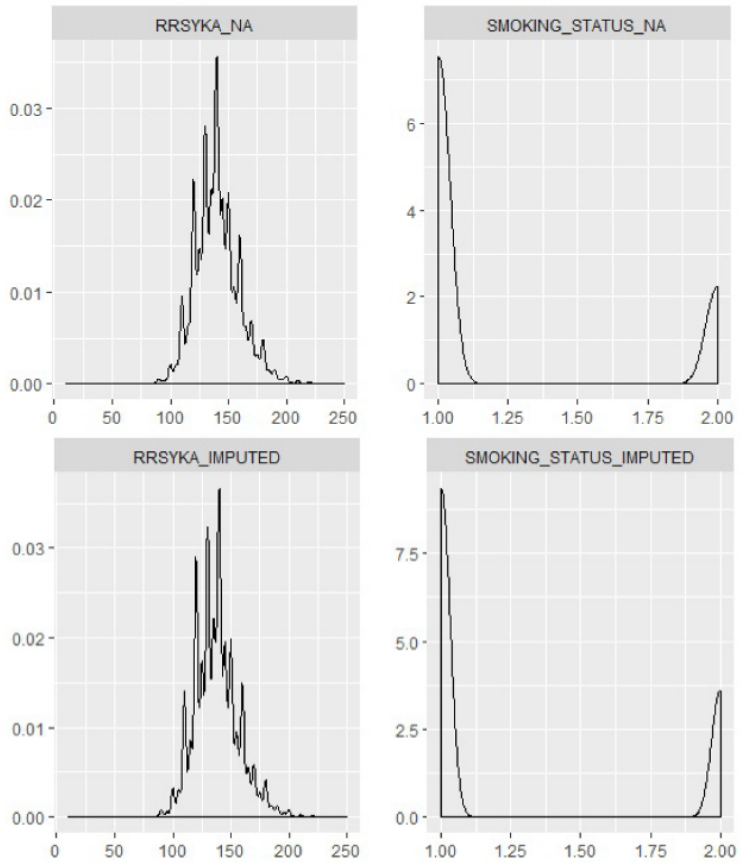


Figure 4.3: Distribution plot comparison of complete case (top) and multiple imputation (bottom)

4.2.3. Data discretization

The data discretization is performed for several continuous variables before doing discrete Bayesian network analysis. The variables that required discretization are age, all the medical measurement variables, as well as several income and wealth-related variables. First, age is split up for every 5-year interval from the earliest age recorded (25 years old) up to the age that is more than 100 years old. Then, the medical measurement is separated by classifying the measurement by consulting an expert and the general practitioner guidelines (NHG, 2019a) of measurement criticality of the values. For type blood pressure and cholesterol related measurement values are separated into 3 - 4 categories, low, normal, high, and critical. While BMI (QUETAO) are separated into 3 categories, normal, overweight and obese (see Table A.14 in appendix A.4). Lastly, for the socioeconomic status such as wealth, income, and assets, the values are separated into three categories of a low, medium, high based on the percentile group of their corresponding values.

4.3. Modelling

Assuming that all the data cleaning and preprocessing are done correctly, now the analysis proceeds to the data modeling. Due to certain computing power limitation in the CBS environment, several assumptions are made related to socioeconomic variables type. First, most of the socioeconomic variables that come from CBS dataset are in the form of percentile group with 100 to 1000 different levels for every variable. Therefore, some of the variables are treated as continuous variables and 4 of the variables the number of levels are reduced into 2 to 3 levels only based on low, normal, and high for income-related variables. As a mat-

ter of fact, treating ordinal categorical variables is a common approach in doing regression analysis (Rhemtulla et al., 2012; Torra et al., 2006; Winship and Mare, 1984) especially if there are only two levels (male and female, above and below the poverty line, etc). However, for ordinal variables with multiple levels, treating them as continuous variables, assumption need to be made regarding the relative difference between consecutive values of the ordinal variable. This means that for some variable, say, percentile group of income variable, that the difference between percentile group of 10 and 11 is comparable to that between 21 and 22. Therefore, as opposed to this assumption, some new variables are created from the percentile group variables.

4.3.1. Kaplan Meier model

The first step of data modeling begins by using the Kaplan-Meier model to give an overview of the survival curve of the dataset. Since the Kaplan-Meier model makes no assumption about the shape of the survival curve, it is wise as the first step of survival analysis to plot the Kaplan-Meier curve to understand the characteristics of the dataset. In this case, the Kaplan-Meier curve is used to gain insight into which time axis is best suited for the analysis. There is two consideration on which variable is suitable to use as the time in the survival analysis since literature suggest (Canchola et al., 2003; Cheung et al., 2003; Lamarca et al., 1998) the possibility of using age as a time instead of following up time that is commonly used. Age as the time scale for survival analysis arguably more appropriate when the purpose of the research is looking at age-specific risk. In addition to that, the inference of risk of higher age group by using age as a time scale is better in comparison with using age as covariate (Lamarca et al., 1998). Then, three different time scale with no covariate are plotted (Figure 4.4) to make a comparison between the different possibility of time scale to be used in the model. As can be seen, using the follow up time as time scale intuitively shows that the probability of surviving from cardiovascular disease event from the time of population registered in the study (1st of January 2011) to the end of study (31st of December 2017) are relatively high with more than 85 percent individuals still alive. In contrast, when age is used as the time scale, the survival curve highlight the higher risk of cardiovascular disease in the different age group. This is more useful in regard to this research since one of the main purpose of the research is also to address the necessity of cardiovascular disease prevention among the younger group of population. In addition to that, two different age variables are also compared. By looking at the survival curve, there is not that much different in terms of the shape of the curve. Therefore, it is decided to use the dynamic age as the time scale as it retains more detailed information at which specific age does a person have an event.

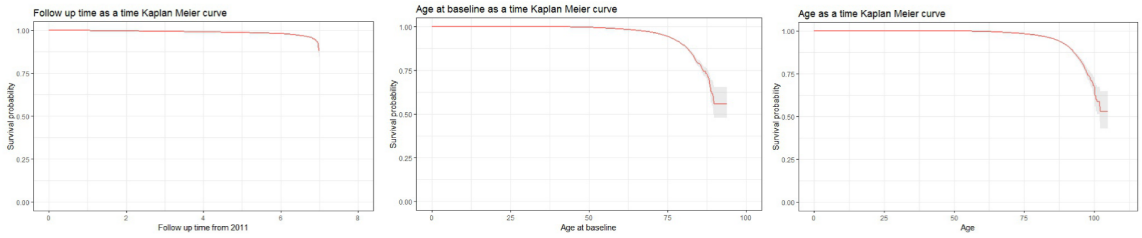


Figure 4.4: Kaplan-meier survival curve of different time scale with first cardiovascular events as the dependant variable

4.3.2. Parametric model

Then, several parametric models are inspected to assessed whether is it possible to fit the dataset into some specific distribution. In this case, the survival curve of the dataset is calculated while also the parameters of the distribution are estimated based on the dataset. Dynamic age is used as the time scale while the first cardiovascular events are used as a dependent variable, and no covariates are used for the model. In addition to that, only survival curve is calculated as it is the fundamental assumption of the parametric models that the survival curve should approximately follow the survival curve of the dataset. Thereafter,

the survival curve of the dataset (black line) and the estimated parametric survival curve (red line) are drawn as can be seen in Figure 4.5. First, the exponential model seems to fail to find the matching parameters to estimate the survival curve of the dataset. This is due to the reason that exponential distribution assumes the natural survival rate that does not vary over time, which represent only λ as the parameter (Table 3.2). Second, the Weibull, Gamma, and Log-logistic distribution seem to work better compared to the exponential distribution in estimating the survival curve of the dataset up to the age of 88 due to an additional parameter that can be optimized. Lastly, the best parametric model that can fit the survival curve of the dataset is considered to be Gompertz, which able to follow the steep slope of the survival curve even at an older age. However, due to the package (flexsurv) and computational power limitation (memory space error) from CBS environment, the Gompertz parametric survival analysis that includes covariates cannot be performed to make the comparison between the Cox-proportional hazards model in the next analysis.

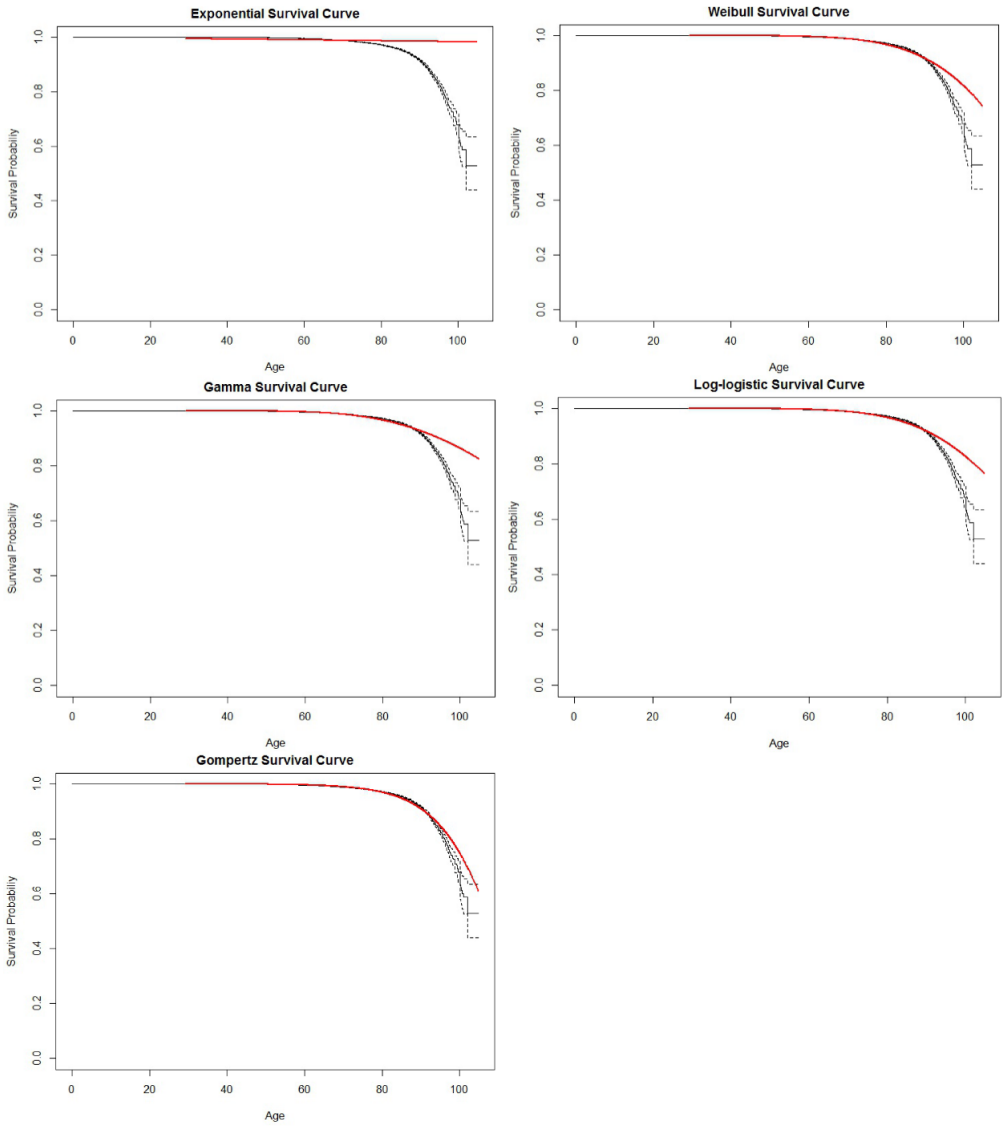


Figure 4.5: Different estimated parametric survival curve (red line) against the non parametric survival curve (black)

4.3.3. Cox proportional hazards model

Since the dataset contains a variable that varies over time, extended Cox proportional hazards model are used in this analysis. By taking into account the repeated measurement of an individual under a small time window, a cox regression is performed separately. Then, the weighted average of all of these cox regression is calculated, resulting in one hazard ratio for the overall analysis for each covariates (Dekker et al., 2008). Zhang et al. (2018) argued that in the presence of time-varying covariates, performing an extended Cox proportional hazards model as opposed to standard Cox regression sometimes can improve the fitness of the model better. However, the next concern of doing Cox proportional hazards with time-varying covariates in longitudinal data is the presence of correlated values in the dataset. This issues actually already been addressed by the author in the R package (Therneau et al., 2013) of "survival" package for the extended cox that used in this research. Though, this applies under the condition that there are no multiple events occurring for an individual. Since the events that used in this research is the first cardiovascular events (death or diagnosis), then there is only one possible recorded event for every individual then it is assumed that the problem for correlated values in this analysis is solved.

For a starter, the variable selection needs to be performed using the Cox model for socioeconomic status variables to reduce the number of variables that are going to be included in the actual model. Due to the uncertainty of missing values in the dataset, the four type of data treatment is used as a way to indicate the consistency of the elimination process of certain variables. The most common approach for the variable selection process is univariate screening, forward selection, stepwise selection, and backward elimination. As for this research, the univariate screening and backward elimination are chosen since literature suggest that on average, these methods produced less bias compared to the other method available (Dunkler et al., 2014). However, for backward elimination of categorical variables, some additional criteria are proposed since the model create multiple dummy variables for each level in categorical variables. The additional criteria being that if at least one of the levels in the categorical variable is significant, the variable is not eliminated from the analysis. In addition to that, the variables that need to be eliminated needs to be at least insignificance in at least three data treatments. These added criteria are inspired by Donders et al. (2006) and Sterne et al. (2009), which provide a guideline to include the different assumption of imputed data into the analysis. The chosen p-value for the univariate screening and backward elimination decided to be higher than 0.25, while variables that have p-value lower than 0.05 are considered to be significant and will stay in the next iteration.

Table 4.2: Cox model iteration summary

Analysis	Blacklisted variables (p-value)			
	Complete cases	Missing indicator	Single Imputation	Multiple imputation
Univariate	INHAHL (0.88)	Vermogen (0.28)	INHAHL (0.98)	INHAHL (0.94)
1st Multivariate	INHAHL (0.79) INHP100HPRIM (0.69) Vermogen (0.57) Welvaart (0.33) INHP100HBRUT (0.27)	INHAHL (0.91) INHP100HPRIM (0.82) Vermogen (0.59) INHBESTINKH (0.31) INHGESTINKH (0.37) Welvaart (0.31)	INHAHL (0.96) INHP100HPRIM (0.54) Vermogen (0.53) INHGESTINKH (0.30)	INHAHL (0.53) INHP100HPRIM (0.59) Vermogen (0.75) INHGESTINKH (0.29) Welvaart (0.35) INHP100HBRUT (0.27)
2nd Multivariate	Welvaart (0.42) INHP100HBRUT (0.30)	INHGESTINKH(0.41) Welvaart (0.39) INHBESTINKH (0.33) INHP100HBRUT (0.25)	INHGESTINKH (0.27)	INHGESTINKH(0.41) Welvaart (0.37) INHBESTINKH (0.33) INHP100HBRUT (0.31)
3rd Multivariate	INHBESTINKH (0.92)	INHBESTINKH (0.58)	INHBESTINKH (0.52)	INHBESTINKH (0.55)

Univariate analysis on socioeconomic factors

The first stage of analysis begins by univariate screening with every single socioeconomic variable used against the dependent variable (time * CVD event) for all four different data treatments. As can be seen in Table 4.2, the INHAHL, and Vermogen variable is considered as insignificance in at least three data treatment and have the potential to be removed in the analysis. However, the literature suggests that removing variables based solely on univariate screening sometimes considered as to be uninformative (Bradburn et al., 2003b;

Dunkler et al., 2014). Since it is possible that even though in the single variable cox model is not significant, it is possible that when the variable model jointly, it will become significant. Therefore, the INHAHL and Vermogen variable will be tested again in the multivariate analysis.

Multivariate analysis on socioeconomic factors

Next, the joint model of socioeconomic status variables is performed. In the first iteration of multivariate Cox analysis, It is revealed that again, INHAHL and Vermogen have p-value more than 0.25, which is insignificance to the dependent variable. And INHP100HPRIM is also insignificance in all missing data treatment while INBESTINKH, INGESTINKH, Welvaart, and INHP100BRUT is insignificance in at least three different missing data treatment. As to give subtlety in eliminating the variables, only INHAHL, Vermogen and INHP100HPRIM are removed for the next iteration. Then, the second iteration of multivariate Cox begins which resulting in the variable that is not significant and not removed in the last iteration to show up again as potential variables that add no variance to the model. Thus, INHGESTINKH, Welvaart, and INHP100HBRUT will be removed in the next iteration since their insignificance is consistent with at least three different data treatment. Afterward, the 3rd multivariate analysis is performed, which resulting in INHBESTINKH to be not significant in all data treatment. After removing all of the variables from all three iterations, the fourth iteration shows no variables that is insignificance except some dummy levels of categorical variables which are not possible to eliminate since the other levels of those categorical variables are significance. Therefore, it is concluded that the socioeconomic variables that add variance to the model areas listed in Table A.11 in Appendix A.3. However, as INHARMLAG with INHARMSOC as well as INHSAMAOW with INHSAMHH have almost the same definition. It is assumed that INHARMLAG and INHARMSOC will perform relatively the same in the multivariate analysis. Therefore INHARMLAG will be chosen as an indicator of income group. Additionally, INHSAMHH will also be used instead of INHSAMAOW as it holds more information which could potentially be meaningful to the interpretation of the effect of household condition to the cardiovascular risk.

Multivariate analysis on socioeconomic factors with medical measurement

Finally, the actual Cox proportional hazards model that includes all aspect of cardiovascular events such as socioeconomic variables, medication history, cholesterol, and blood pressure measurements, as well as a smoking status variable, can be modeled (Table A.12 in appendix A.3). However, as medical measurement data has at least more than 67 percent missing, the assumptions for using complete cases and the missing indicator will not hold since the sample size of each different data treatment will vary by almost half of the dataset. Therefore, as to minimize all the assumption about on which dataset that the conclusion of the modeling result is withdrawn, only two datasets are considered in this case, single imputation that used for the socioeconomic status variables while all medical measurement used multiple imputation values. The reason is that it is assumed that the single imputation that is performed for socioeconomic status variables have smaller assumption (drawn for mean or mode of an individual) compared to the multiple imputations. However, for the medical measurement as the portion of missing is really large, more assumption is needed to fill in the missing values; thus, multiple imputations are chosen for the final model. Fortunately, there is no missing value in medication and treatment history since it is assumed that if there is a recorded cardiovascular-related drug in that individual years, it will be labeled as one which translates to ”have cardiovascular medication”. The recorded how much medication of a person takes every year also used as a variable in the final model. Nevertheless, this research makes no distinction between cholesterol-lowering drugs and blood pressure-lowering drugs since there is knowledge limitation about some specific ATC code that is used in the dataset. Two iterations are performed in the actual model with all the important variables with the first one being that the variables are put as it is while the second iteration focuses more on solving issues of the violation of the proportional hazards in Cox regression. The detailed explanation about this will be discussed in the chapter 5 in section 5.1.

4.3.4. Discrete Bayesian network model

The discrete Bayesian network model in this research is used as an experimental modeling technique to put more assumption into the model structure. The main paper that used as the main reference for this model is written by Bandyopadhyay et al. (2015) and Scutari and Denis (2014) which describe the different method of doing Bayesian network for time-to-event data. This research chooses one method of doing a Bayesian network, which assumes that the repeated measurement of an individual as a different person, and there is no temporal assumption from the data. This implies that for every recorded data, each variable is assumed to only contribute to the dependent variable of that particular time when the data is recorded. As a result, the number of individuals that previously recorded to be 43,949 will become 377,885. These assumptions are made since to add a temporal assumption in the model; the data need to be transposed back into the "wide" structure which will result in the increase of the number of columns by N number of variables times the K number of unique repeated measurement. In this case, 20 variables that used in the model, need to be separated into each recorded time, for example, 7 (assumed yearly record, while in actual data is mostly monthly even daily), then there are 20 * 7 years which is 140 columns. The second problem is related to how to create the model structure in the Bayesian network. In Bayesian network with the temporal assumption, it is assumed that the dependence of variable in the future, should only depend on the present. For example, if a person takes a cardiovascular medication in the year 2011, it is assumed to affect the low blood pressure measurement in the year 2012 and not the accumulated effect form 2007 to 2011. Although conceptually, this is made much more sense, to be able to model this interaction across time makes the model structure much more complicated and might not improve the model result significantly (Server, 2017). Therefore, for a starter, the current assumptions about the repeated measurement of an individual are used as bases of the analysis. However, not every socioeconomic variable can be used as an input for discrete Bayesian network, the variables that are going to be considered in this model are reduced to the only variable that is categorical or equivalent of the categorical values (see Table A.13 in Appendix A.4).

Structure learning

It is common practice when performing discrete Bayesian Network to rely on expert or prior knowledge of the system of interest and use that information to define the model structure (Scutari and Denis, 2014). This type of modeling practice that used Bayesian Network as a *expert system*, will also be used in this analysis. There are two steps in which the model structure are decided. First, the hill-climbing algorithm is performed with the dataset, and then the expert is asked related to some dependencies and independencies that is produced by the hill-climbing algorithm. As there might be several pieces of information that the expert might not know, some literature is also consulted. Fortunately, there are several literatures that already studied in-depth about different variables dependencies in cardiovascular disease topic. First, Bandyopadhyay et al. (2015) used clinical knowledge of an expert to define a model structure. Suggests to model commodity (Whether or not a patient had experience cardiovascular disease), age, and sex as the first-order relationships with other medical measured variables and medication. While, the cardiovascular events are dependence on every node that exists in their model. On the other hand, Twardy et al. (2004,0,0) compare different literature and different cardiovascular risk model (ie. CUORE and PROCAM) to gain insight about the plausible cardiovascular model for Bayesian Network. The same modeling structure also found in this literature, with sex as the first-order variable while in this case age is considered in the second-order along with the other type of measurements. However, all of the literature above did not mention some socioeconomic status variables as part of their analysis. Therefore, the model dependencies for the socioeconomic status variable are assumed by the author. There are two main assumptions for the socioeconomic status variables, first ethnicity, poverty as well as prosperity connected to smoking status (Miner et al., 2014) while the only cardiovascular event is directly dependent on poverty variables. The summary of the model structure can be seen in Figure 4.6 with all variables that are colored modeled as one single cluster that consists of multiple node. Thus, if at least one node from this cluster is connected to the other node, it is assume that the whole node from that par-

ticular cluster is connected to the end node as well. This cluster of node is one of the way to represent the arc connection between independent covariates and dependent covariate (first cardiovascular disease event). Since it is impossible to connect every independent covariates to the dependant covariate. For more detailed model representation, please refer to Figure A.2 in Appendix A.4.

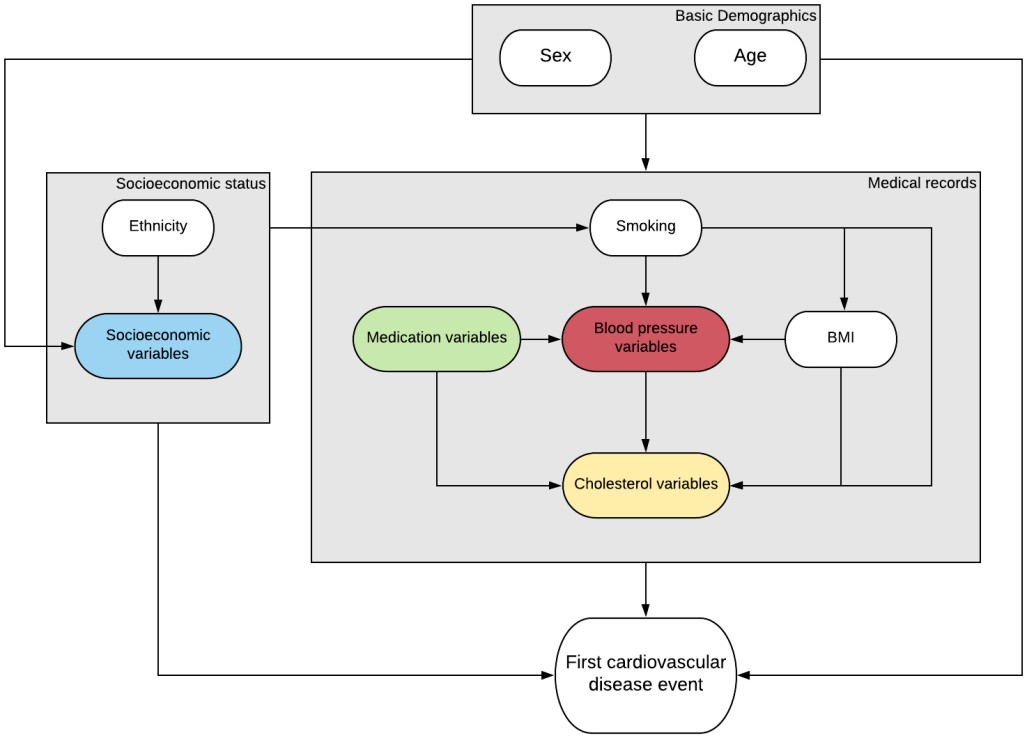


Figure 4.6: Simplified version of Directed Acrylic Graph of Discrete Bayesian Model

Parameter learning

After the model defined, the global distribution of each unique combination of events needs to be calculated. In the case of Discrete Bayesian Network, these joint probability distributions are represented as multinomial distribution. As mentioned in Chapter 3, there are two ways of calculating the estimated probability of each unique combinations of events in Discrete Bayesian Network, Maximum likelihood estimation, and Bayesian approach. Both of the methods are used in this modeling practices. While Maximum Likelihood Estimation is quite straight forward to implement, Bayesian Network is not. It requires additional argument, imaginary sample size (iss) to be put into the function. Literature suggests that a good rule of thumb is to use a number that less than 20 as the iss value (Scutari and Denis, 2014). The higher the iss value, the more uniformly distributed the posterior distribution is going to be. Therefore, 10 are chosen as the iss value to allow some small fraction of prior distribution to affect the posterior distribution of the parameters of the multinomial distribution.

Bayesian Inference

In Bayesian Network, the inference of the modeling result can be done by *querying* the conditional probabilities of the random event given the evidence. For example, it is possible to put the query of the conditional probability of having a cardiovascular event given the evidence that the population is male. These type of queries can be done in two ways, the exact inference, and the approximate inference. The exact inference will use the events and evidence to get the parameter that already calculated in the model while approximate inference works by running some Monte Carlo simulations to replicate the dataset given the defined

distribution of the model. This research used the approximate inference with 1 million sample replication as it is incredibly versatile in terms of the conditional probability results. In addition, only some variables that are significant in the Cox proportional hazards model are produced for the output to ensure the consistency in drawing the conclusion from the result. The inference is performed by putting the Bayesian events as the combination of nodes with cardiovascular events variable that consist of two levels of either ”Events” or ”Censored” and the other variables of interest. These variable of interest are one variable from either socioeconomic status or medical measurement with/without the age variable (which will be used as a time axis to compare with cox). Since this research is particularly interested in the overall risks of the population, the Bayesian evidence that is used is the population that either is ”Male” or ”Female” in the Sex variable. As a result, the frequency table in which all the nodes are specified are produced, then the probability table is derived out of it. Since what this research also interested in hazard and survival function of different group of population, by following the definition of Function 3.7 then convert it into Function 3.5 and Function 3.6, the probability table are calculated vertically for every unique group (ratio of people who have ”Event” and total population at risk at that particular group).

5 Results

5.1. Model Validation and Verification

The models are validated and verified in two ways, logical reasoning from an expert along with the literature and statistical test of the model. The former definition is quite straightforward; the expert is consulted about the data modeling process as well as the modeling result. In addition, statistical tests are performed by checking the residuals of the cox modeling result that constructed with the standard Schoenfeld residuals tests. This residual test is performed to test whether the proportional hazards, which is the main assumption of the Cox proportional hazards, are violated or not in the model. As mentioned by Kalbfleisch and Prentice (1981) and Uno et al. (2014) "When the PH assumption is violated (i.e., the true hazard ratio is changing over time), the parameter actually being estimated by the Cox procedure may not be a meaningful measure of the between-group difference; it is not, for example, simply an average of the true hazard ratio over time". Therefore, checking the proportional hazards assumption of the Cox model can be considered as the quality and reliability check of the modeling result. Moreover, the validation of a discrete Bayesian network is done by testing the interpretation of Bayesian Network model against the Cox proportional hazards model interpretation and check whether both models have reached the same conclusion.

5.1.1. Cox proportional hazards model validation

There are two methods of validating whether or not the model is violating cox assumptions, graphical diagnostics, and statistical test. Graphical diagnostics for the Cox proportional hazards model are not going to be considered as model validation because of two reason, first it is highly subjective to the researcher to determine whether the residual has a pattern against time; and secondly, even if the graphical diagnostics are performed, the residuals plot that is produced cannot be shown in this report as the CBS environment (server) forbid such plot to be published outside. Thus, one of the most common residuals test called the Schoenfeld Residuals Test is preferred in this research. Schoenfeld residuals test are used to assess the trend against time or lack of proportionality. If there are time-dependent covariates, this will result in generalized linear regression on functions of time to produce a non-zero slope, which is an indication of a violation of proportional hazards assumption (Abeysekera and Sooriyarachchi, 2009).

The first attempt in performing multivariate analysis with medical measurement shows that there are several covariates that violate the proportional hazards assumption which indicated by the Schoenfeld residuals p-value lower than 0.05. This p-value shows that the covariates are significantly contributing to the non-zero slope in the residuals plot. The variables that is violates the proportional hazards assumptions are (check Table B.4 in Appendix B.2), MEDICATION COUNT, INHBBIHJ (22, 30), INHP100HBEST, VEHP100WELVAART, Armoedegrens ,SMOKING STATUS ,MEDICATION LABEL, Herkomst gehercodeerd. As a result, additional steps are required to solve the proportional hazards issues. There are at least three method to deals with proportional hazards violation in Cox model, stratification (Ata

and Sözer, 2007; van Houwelingen and Eilers, 2000), step function (Therneau et al., 2013), and adding sophistication of the model by adding interaction with time (Ata and Sözer, 2007).

In stratification, the hazards model (function) calculation is separated for each stratum (levels) of p covariates that do not satisfy the cox proportion hazards assumption. In this research, variable Herkomst gehercodeerd (ethnicity) is one of the examples of the use of stratification. This method is desirable because of two reason, the variables that are used are categorical variable, and there is a possibility that the assumption is violated due to the huge difference in sample size. However, the disadvantage of using stratification is that the variable that is chosen to be stratified will lose its function as an explanatory variable in the multivariate analysis. In fact, the dataset that is used will be separated based on levels of the stratified covariates then the regression coefficients will be estimated by multiplying the likelihood function of each stratum (van Houwelingen and Eilers, 2000). On the other hand, step function works by calculating different coefficients according to the time intervals when the proportional hazards are violated. For example, if the Schoenfeld residuals plot shows that covariates p have a non-zero slope at time $T > 50$, then the covariates p will be separated into two different covariates $p_{T<50}$ and $p_{T>50}$ (Therneau et al., 2013). However, this method requires graphical diagnostics, which is currently not possible to report if it were used. Additionally, when the Schoenfeld residuals plot is inspected, there is subtlety in the change of slope in all of violating variables which hard to deduct in which time period that the variables start to violate the proportional hazards assumptions. Therefore, the step function technique is not used in this research. Lastly, there is a possibility to add interaction to the violating covariates with time. As a matter of fact, this technique is the most efficient way of solving the violation of the proportional hazards since it is not only identified in which period the violation of the covariates starts while forcing the covariates to goes back into its constant trajectory of the hazard ratio (Ata and Sözer, 2007). Therefore, the variables besides Herkomst gehercodeerd (ethnicity) used additional interaction with time. However, using age as the interaction term actually pose several issues in the algorithm that is used; thus age at baseline (2011) are used instead of the dynamic age that used as a time axis. The result of second residuals test can be seen in Appendix B.2 Table B.6. It is worth to mention that the likelihood ratio test for the Cox model with interaction since to hold higher value compared to the one without the time interaction, although it is not significantly higher. This means that in terms of performance, both models could yield the same quality, although, in the second model, there is a violation of the proportional hazards assumptions. Thus, to make an interpretation, only the second model is going to be interpreted.

5.1.2. Discrete Bayesian Network model validation

Before comparing the different interpretation of Cox and Bayesian, it is conceptually possible for Bayesian to replicate the Kaplan-Meier curve with age as of time axis. In case of Discrete Bayesian Network, to be able to replicate the survival curve, the hazard rate of each age group are converted into its $exp(-\Lambda)$ form. As there are two methods of calculating the parameter in Bayesian Network, only the maximum likelihood estimation are compared with the Kaplan Meier survival curve. Kaplan-Meier is used as the first indicator of survival analysis in Bayesian Network and can be seen in Figure 5.1.

Given the current assumption of Discrete Bayesian Network that tries to assume that each repeated measurement of individuals that are censored are a separate subject. This is resulting in underestimation of the hazard rate of people who have a low number of people count and a low number of a cardiovascular event. In this case, the people who have the age above 90 are the example of such a case. However, this check does not mean that the result of Discrete Bayesian Network result is invalid. This test has proven that the result from low sample size with a low number of events should be interpreted carefully and should not be underestimated. Therefore, for the next validation part, only variables that have a low number of levels are considered for interpretation and comparison between the two modeling approaches.

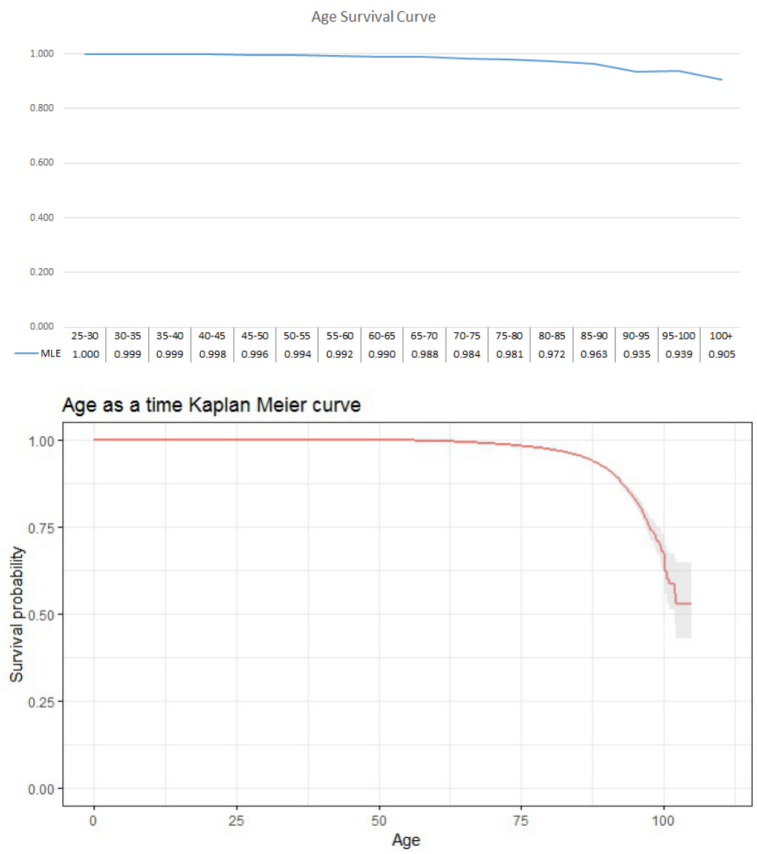


Figure 5.1: Discrete Bayesian Network (Top) comparison with Kaplan-Meier survival curve (Bottom)

5.2. Model Interpretation

As the number of variables that used in the model is quite high, all the significant results are placed in the Appendix B.2 for Cox proportional hazards modeling results while Appendix B.3 listed the results of Discrete Bayesian Network model. Therefore this section will only highlight several interesting modeling result in both cases. Please also note that the author is not interpreting the result in terms of the ”real value” of the coefficients but instead only refer to the generic interpretation of the model such as high, low, higher, lower, good, bad, etc. The reason for this is due to model and data limitation that exist in this research which will be explained in the next chapter.

5.2.1. Cox modelling result

The interpretation for the Cox model is performed only for the variable that is significance or p-value lower than 0.05 as the standard in statistical analysis. This is because there is a high probability of mistakenly reject the effect of covariates effect on the hazards rate if there high p-value. Therefore to ensure that to minimize the possibility of make the wrong interpretation of the result, only covariates with a p-value lower than 0.05 are interpreted. In addition to that, there is two main interpretation when using Cox model, the good prognostic which represented by the hazard rate (Exp(coef)) lower than 1 and bad prognostic for the opposite. Good prognostic refer to particular covariates have an impact on the lower probability of subjects experiencing the first cardiovascular event and vice versa for bad prognostic.

First, Holding the other value constant, having higher INHARMLAG (income according to the poverty line) relate to good prognostic. This means that the higher the individual percentile group away from the poverty line, the less likely that the individual to experience the first cardiovascular event. On the other hand, if INHARMLAG separated only by two groups,

above and below poverty line which is represented by variable Armoedegrens the result shows that in the older age, people who live above the poverty line have proven to have lower risk of having a cardiovascular disease event compared to the one that is living below the poverty line. On the other hand, VEHP100WELVAART (a measure of prosperity or welfare) also shows the same result as INHARMLAG with good prognostic for people with higher prosperity percentile group. Second, holding the other value constant, having GBAGESLACHT is 2 (being female), relate to good prognostic by factor 0.6569. This means that female have a lower risk of having cardiovascular disease events compared to men. Third, the result also shows that having cardiovascular medication is more beneficial to older people compared to younger age group. This indicated by the variable MEDICATION LABEL interaction with AGE AT BASELINE have hazard rate below than 1. Fourth, individuals that belong to INHSAMHH 12 (Single man from pension age) as opposed to INHSAMHH 11 (Single man to pension age) which is used as a reference of the dummy variable have the hazard coefficients of 0.7146 which related to good prognostics for covariate INHSAMHH 12. This means that a single man that is older have better survival compared to a single man that is younger. Additionally, it is also identified that people who have income from property (INHBBIHJ 30) have better prognostics compared to people who have income from salary (INHBBIHJ 11) with hazards coefficients of 0.117. Finally, although the stratified covariate cannot be observed in the result of regression covariates, the likelihood function of each stratum can be observed through graphical representation of the survival function as can be seen in Figure 5.2. From this graph, it can be said that the median survival time for every ethnicity given the current dataset was met after 75 years of living. This means that more than 50 percent of people in the dataset have their first cardiovascular event starting from the age of 75 years old.

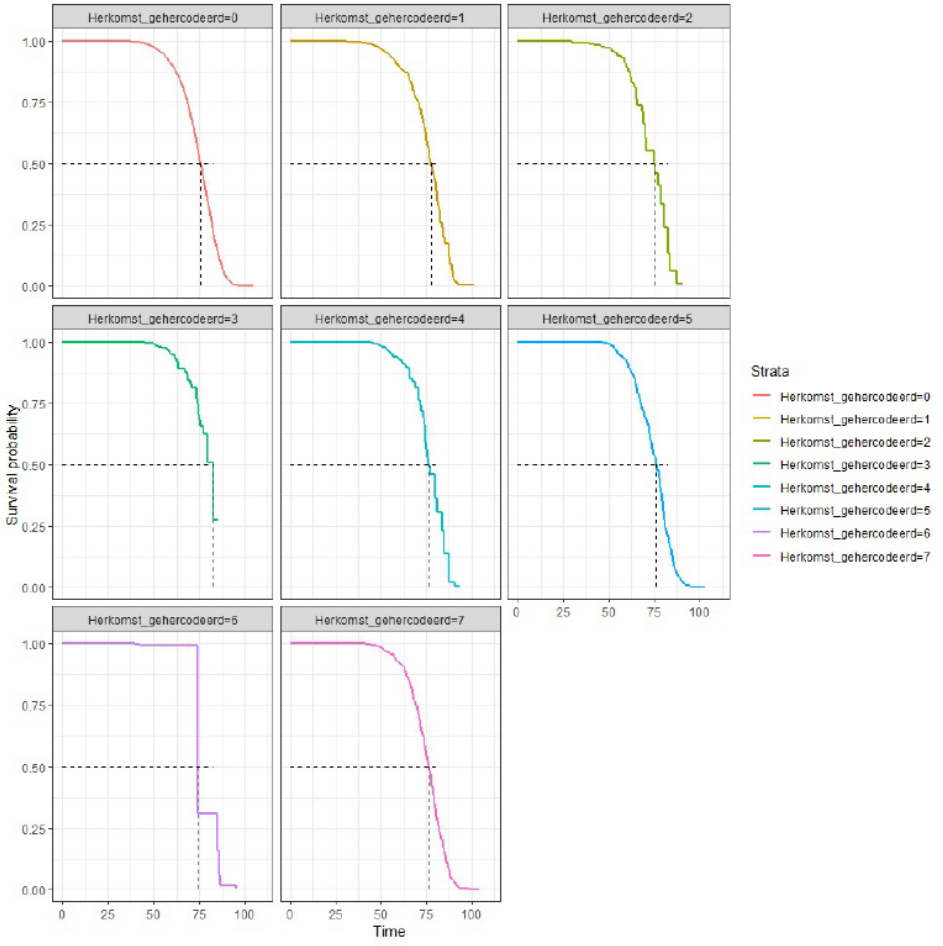


Figure 5.2: Cox model for different ethnicity stratum

5.2.2. Discrete Bayesian Network model result

Interestingly, the same result can be seen for the socioeconomic status that is related to the poverty line (equivalent with Armoedegrens) and income (equivalent with Gest Besteedbar Inkomen) in Discrete Bayesian Network. As can be seen in Figure 5.3, the worse the condition of poverty and income, the higher the risk of having cardiovascular events. This result is similar to the Cox modeling result of Armoedegrens and INHARMLAG that have Exp(coef) or hazard rate less than 1, which lead to good prognostic in higher values of these variables. This result also consistent in two different parameter learning method (maximum likelihood estimation (MLE) and Bayesian approach) although the value of the hazard for Bayesian is bigger compared to MLE. This is due to the fact that there is an allocation of prior probability distribution in the posterior probability estimation.

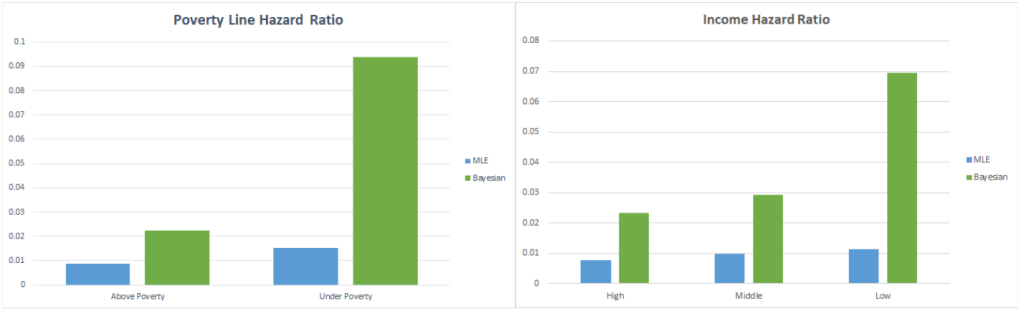


Figure 5.3: Discrete Bayesian Network result for poverty (left) and income (right) covariates

However, for covariate prosperity, Discrete Bayesian Network shows a nonlinear relationship from low to high prosperity. And also there is a result inconsistency between two different parameter learning approach for prosperity covariate. As can be seen in Figure 5.4, calculation of the parameter by using MLE shows that people with high prosperity, have the highest likelihood to have a cardiovascular event while people that belong to the low and medium prosperity have a comparatively lower probability of having an event although the difference between each prosperity category is comparatively small with each other. On the contrary, The result using Bayesian approach shows quite the opposite result. The individual that belongs to the low prosperity now considered as the one that has the highest risk of having a cardiovascular disease event. While the individual that belongs to high and medium have a roughly lower risk of having an event of interest, this inconsistency actually gives additional information about the reason why in Cox proportional hazards model VEHP100WELVAART variable was violating the proportional hazards assumption. On the other hand, the same result also produced for variable GBAGESLACHT, with higher risk among men compared to male with both parameter learning method producing the same result (Figure 5.4).

Unlike the Cox proportional hazards model which need to used stratification to model variable ethnicity, in Discrete Bayesian Network, the hazard rate of a different ethnic group can be observed directly. Likewise, the contradictory result is produced in both methods of parameter learning as it can be seen in Figure 5.5. Conceptually, if it was possible in Cox proportional hazards to use ethnicity without stratification, the interpretation of Cox modeling result for ethnicity variable would somehow match with the result of MLE in Discrete Bayesian Network. Then, by taking that into account, the conclusion of the MLE would be that people that have Turkish as their ethnicity are the most prone to having cardiovascular disease event while polish people are the least susceptible ethnicity in having the event of interest. However, if the Bayesian approach is used instead of MLE, the conclusion will show that the polish people are the most prone individual to cardiovascular disease, which is exactly the opposite result in MLE. This result also produces some additional insight into how well the Bayesian approach deals with a different number in sample size. In MLE, there is a likelihood that some ethnic groups that contain a bigger sample size (overall population and

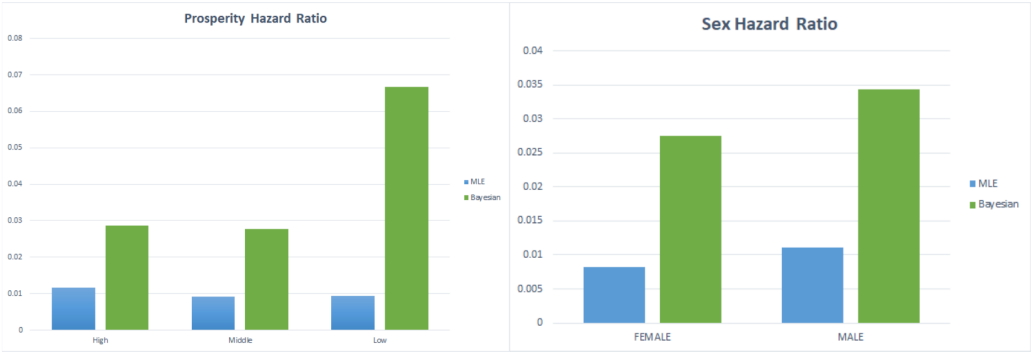


Figure 5.4: Discrete Bayesian Network result for prosperity (left) and sex (right) covariate

event count) will have a higher risk of experiencing an event. While in Bayesian approach due to the weight that put over the prior probability distribution, there is an equal chance of ethnic group with small sample sizes with low event rate to have the same chance as the having higher probability as the ethnic group that has high sample sizes.

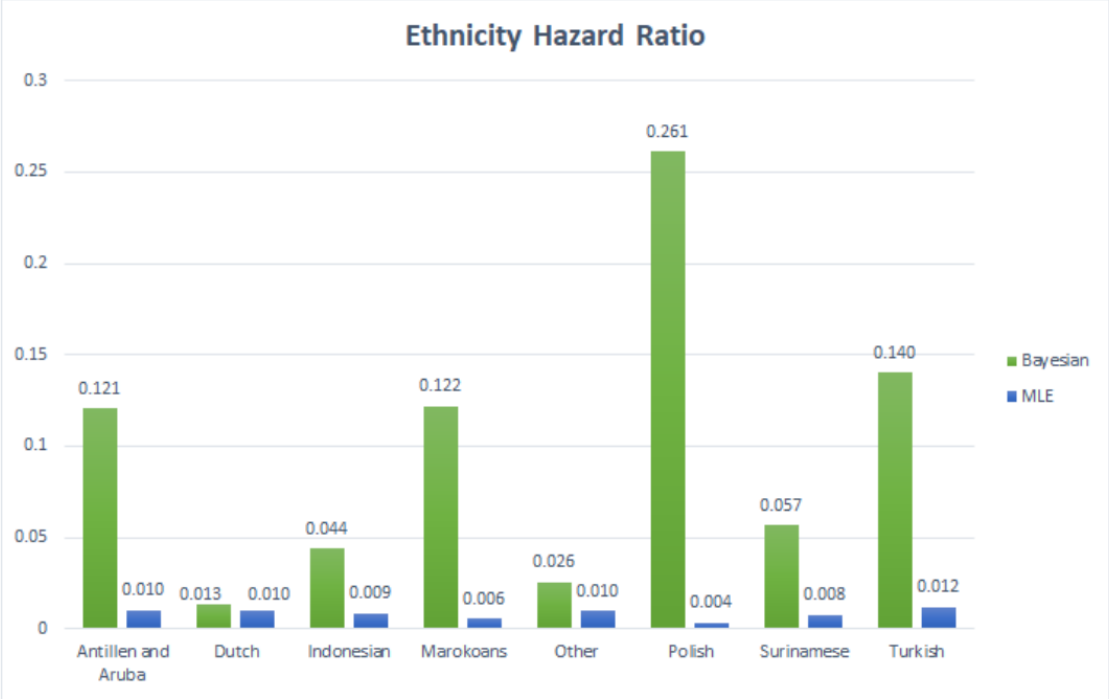


Figure 5.5: Discrete Bayesian Network result for ethnicity covariate

6

Conclusion and Discussion

6.1. Conclusion

In summary, there are four different modeling approach that explored in this research, the non-parametric model with Kaplan-Meier survival model that used to identify the nature of the survival curve in the dataset that used. Then, several parametric models are fitted into the dataset to check the possibility to use distribution assumptions about the survival curve. Subsequently, the Cox proportional hazards models are modeled for backward elimination of socioeconomic status variable, which then used in the final model with medical measurement records. Next, Discrete Bayesian Network is used to add perspective in other modeling approaches as opposed to the Cox proportional hazards model to draw the conclusion. Finally, Cox proportional hazards model is validated with Schoenfeld residuals test for checking the proportional hazards assumption while Discrete Bayesian Network used the Cox model as the standard of model validation. As a result, conclusions can be made based on the two sub research questions and one main research question that is proposed in this research.

- Sub research question 1 -

"What are socioeconomic status that is feasible to include in the cardiovascular risk model?"

Based on the backward elimination that is performed in this research, there are at least 12 variables that are significant to the survival times of dataset population. And these 12 variables are found to be consistent regardless of 4 different imputation method that each have different missing values assumptions. However, after removing two covariates (INHARMSOC and INHSAMAOW), adding medical measurements covariates into the multivariate analysis of Cox proportional hazards model and interaction with time (age) for some covariates that violate the proportional hazards assumption, only seven covariates are proven to be significance to the survival times. These variables are INHARMLAG, INHPOPIIV, INHSAMHH, INHBBIHJ, Armoedegrens, VEHP100WELVAART, and Herkomst gehercodeerd. While the other three variables (INPPOSHHK, Gest Besteedbar Inkomen, and INHP100BEST) failed to be significance after adding the medical measurement variables into the Cox proportional hazards model. Furthermore, in all seven significant variables, not all levels in categorical variables are considered significant. For instances, INHPOPIIV only has categorical 2, 3 and 8 are significant; INHSAMHH only 12, 55 and 71; INHBBIHJ only 26 and 30 that are significant; while Herkomst gehercodeerd only categorical variable 3 are significant.

There is no denying that across all the socioeconomic status that tested in the risk prediction model, the lifetime financial condition of a person has a higher correlation to the early exposure of cardiovascular diseases compared to the other socioeconomic status. Variables

such as income (INHARMLAG), welfare (VEHP100WELVAART), and poverty line (Armoedegrens) are proven to be more significant compared to the other covariates. This significancy does not only comes from the statistical test that performed in this research, but also from the interpretation and intuitive reasoning. Firstly, in terms of the modeling process and result, income-based socioeconomic status is relatively easier to model since most of the variable, there can be considered as an ordinal variable. The result is also much easier to communicate since the reference point for the result interpretation are usually taken from either the lowest or highest values, which assumed to be linear as the values go up or down. Secondly, the financial condition of an individual that is correlated with the event of interest can be explained in many different ways as opposed to the other covariates. Therefore, it is concluded that one of socioeconomic status that is worth to be included in the cardiovascular risk prediction model is definitely the measure of **financial condition (income, wealth and/or assets)** of an individual. However, other types of socioeconomic status variables that could also prove to be beneficial to be included in the risk prediction model such as **martial status (household composition)**, **source of income**, and **ethnicity** could also be beneficial in this case. As can be seen in the interpretation of single man (INHSAMHH 12 and 11) and income from property (INHBBIHJ 30 and 11) and also Herkomst gehercodeerd (or ethnicity in Discrete Bayesian Network model).

- Sub research question 2 -

"What are the trade-off between the various model that is used for cardiovascular risk modelling?"

As briefly mentioned in Chapter 3 for statistical method survival analysis, each model has different advantages and disadvantages. However, after implementing each model carefully given the dataset that used, there are several additional insight that can be learned from each different models. Firstly, it is not exactly true that multivariate analysis can not be performed using the Kaplan-Meier model. It is possible to compare different covariates effects against the survival times in Kaplan-Meier estimate, but to interpret the result from it would be extremely hard given that there are multiple values inside the covariates. For example, if two covariates Sex (with two levels) and ethnicity (with eight levels) are used in the Kaplan-Meier model, there will be 16 different modeling result for every combination of unique values in all the parameters. However, if the number of covariates and the levels that are going to be used is low, it is encouraged to use Kaplan-Meier model as the assumptions are more relaxed compared to the other type of modeling. In addition to that, the Kaplan-Meier curve can also be used as an indicator for checking whether some values inside a variable can potentially violate the Cox proportional hazards assumption. By plotting the survival curve of each categorical covariates one by one, the researcher can observe whether there is a non-proportional line or nonparallel line, which is the indication of non-proportional hazards. Therefore, it is also recommended to use this model as the preliminary steps to observed the natural survival time of the dataset.

Secondly, for the parametric model, the literature stated that it is relatively easier to perform the analysis compared to semi-parametric. This could possibly be true in this research if it was possible to perform the Gompertz parametric model in the CBS environment. In the parametric model, the only assumption that the researcher needs to make is that the survival times of the dataset need to follow a certain type of distribution. If these assumptions are met, then multivariate analysis can be performed without some additional check for the covariates. On the other hand, the semi-parametric model does not assume some distribution about the survival times, but the effect of every single covariate needs to be checked whether the hazards rate are proportional or not. This drawback is the main reason why in the literature, it is mentioned that the parametric model is relatively easier to implement compared to Cox proportional hazards. However, due to strong assumptions of the survival times, Cox proportional hazards model tends to be more favorable compared to the parametric models. Overall, both parametric and semi-parametric model is really useful as it can accommodate the multiple variables in the analysis. Although the assumptions of both models are funda-

mentally different, the result that drawn from both analysis could potentially be the same if the same dataset is used and the assumption for each model are employed correctly.

Besides those three statistical methods, a machine learning-based method is also used in this research through Discrete Bayesian Network. In general, there are three main comparisons between statistical method and machine learning method which are related to censoring assumptions, sample size assumption; and intuitive model and interpretation. In all statistical method, the number of people at risk at time *t* is always depended on the number of people that experiencing an event at time *t-1* and censored individual at time *t-1*. Therefore, exclusion and inclusion of individual that is censored can easily be done in standard survival analysis. However, in Bayesian Network exclusion of censored individual at *t-1* is conceptually harder to imagine compared to the statistical method. In this research, the assumptions to include censored individual happens only in the overall hazard ratio (no time axis) while if there is time (age) survival analysis, the inclusion and exclusion of censored individual failed to replicate the Kaplan-Meier survival curve (see model validation for Discrete Bayesian Network). This proves that in Discrete Bayesian Network, exclusion and inclusion of the censored individuals across time is much harder to do compared to the statistical method. On the other hands, the statistical method relies quite heavily on the sample size. The larger the sample size, the better the prediction will be while the smaller the sample size, the less accurate the prediction will be, and the more likely that over-fitting will occur. In contrast, there is an option in Discrete Bayesian Network to be able to calculate the parameter of the model by using the Bayesian approach, which can allocate a prior distribution over the posterior distribution. By using this method, the unequal distribution of sample size over some sample group will somehow be equalized since there is a possibility of the posterior distribution to be marginalized by the prior distribution. Lastly, modeling in Bayesian Network is much more intuitive in terms of building the model structure as well as interpreting the result. Through directed acrylic graph, it is clear and transparent, which covariates dependence on which covariates while in multivariate analysis the confounding factors of different covariates are more subtle and in some cases are hidden. In addition to that, due to the possibility of using Monte Carlo simulation in the R packages (bnlearn) that used, is it possible to synthetically recreate the dataset given the event and evidence that set in inference function, thus making the model interpretation of Discrete Bayesian Network is relatively easy compared to Cox proportional hazards. For example, in the inference function, if the parameter of the model structure is already known, by just specifying ”the events of interest” and the ”evidence”, the probability of that event happening can be directly inferred. For example, if researcher want to know what is the likelihood of ”Male” population to have ”the first cardiovascular event (Event)” then by specifying the ”evidence” as ”Male” and ”the events of interest” as ”the first cardiovascular event (Event)” then the probability of those ”event” happening will be produced. And it is also possible to add multiple conditions for the ”evidence” and ”the events of interest” which then makes it easier to make interpretation compared to Cox proportional hazards that need to calculate the hazards ratio based on the coefficients that produced by the regression. Cox modeling result interpretation is even harder when there is an interaction with the covariates.

- Main research question -

How are different socioeconomic status contribute to cardiovascular diseases risk of the Hague population?

In summary, there is definitely added value in including socioeconomic status to the cardiovascular diseases risk model. First, this research identified that people with higher income and people who live further away from the poverty line are definitely more fortunate since they have lower tendency to experience the first cardiovascular events, especially in older days. Secondly, for welfare or prosperity, which is a combination of income and wealth (financial assets), as long as Cox modelling and Discrete Bayesian Network (with Bayesian approach for parameter learning) result considered, it is indeed that people with lower prosperity are considered to be more prone to cardiovascular diseases compared to people with higher pros-

perity. Thirdly, there are interesting findings related to marital status, especially for a single man. As it is confirmed by the model, a man, in general, have a higher risk of having a cardiovascular disease compared to female. By combining that fact into the marital status and pension age, this research finds out that there is a better prognostic for single man from pension age compared to a single man to pension age. This means that the younger male population that is single are relatively more prone to cardiovascular diseases compared to a single man in the older population. Three things need to be considered when interpreting this result, the pension age in the Netherlands is 65 years old, while also looking at survival curve of the cox model (Figure 5.2) which the curve start decline at the age of 50 and median survival curve identified at the age of 75. Then one can infer that there are more cardiovascular events happenings for a single man with age 50 to 65 years old compared to 65 years old above. Fourth, it is also discovered that the source of income can also play a role in identifying which population group is more prone to cardiovascular disease. In this case, the modeling result only identifies that people who have income from property have relatively better survival rate compared to people who earn income through salary. Lastly, ethnicity could also contribute to the early prediction of cardiovascular risk as can be seen in the result of both models (Cox and especially Bayesian Network). Although in the Cox proportional hazards model, the ethnicity is modeled as stratification of the dataset. Excluding this variable could potentially create more covariates violating the cox proportional hazards assumptions. However, in Discrete Bayesian Network, it can be seen clearly which population group are more susceptible compared with the other. In this research, Turkish and Polish people are considered as the most susceptible population group compared to all sub-ethnic group.

In spite of the points mentioned above, when considering to include socioeconomic status in the cardiovascular risk prediction model, one should not overestimate the result that comes from the model. The use of socioeconomic status should just be used as a preliminary identification of the individual that is susceptible to cardiovascular diseases. If it is indeed the case that individual belongs to the susceptible population based on their socioeconomic status, then the practitioner should monitor those individual while closely also encouraging them to take some necessary medical measurement. Thus the use of socioeconomic status variables is therefore primarily as means for the practitioner to encourage patient to take the medical measurement which then can be used as the primary indicator to make the judgment for cardiovascular diseases prevention.

6.2. Social relevance and recommendations

This section will try to explain the implication of the modeling results related to the decision making process among the street-level bureaucrats. In chapter 2, it is explained that general practitioner, according to the nature of their work, could be considered as street-level bureaucrats. However, Lipsky (2010) also mentioned that there is some specific behavior that street-level bureaucrats employed given the nature of their work. Erasmus (2015) summaries of street-level bureaucrats behavior will be used in constructing the recommendation to the general practitioners. Thus, by reflecting the modeling results respected to the street-level bureaucrats, this research will try to explain the advantages of the modeling results to inform the general practitioner as a decision-maker in cardiovascular disease prevention.

Services rationing as a means to manage and conserve the street-level bureaucrats' resources

Due to high demand and shortage of medical practitioner in the Netherlands, it is not uncommon that they are rationing the services provided to the patients. This research, in particular, could help general practitioners in rationing their services to more specific cardiovascular diseases susceptible group of population. Rationing, in this case, means to allocate the general practitioners resources (time, expertise, and care) to the one who needs the most care. According to the modeling results, there is at least two crucial information that general practitioners can use in rationing their services. Firstly, all income-related statement from a patient can be used to leverage the way of general practitioners prioritizing or allocating their time. For example, if general practitioners can get information about their patient concerning how much money they make per some period (i.e. monthly). Then, if this particular patient is considered to have income below the poverty line, general practitioners could prioritize these patients more, compared to the patient who lives above the poverty line. Additionally, in terms of simplicity of the socioeconomic variables, poverty line found to be the most convenient one compared to the other variables. Questionnaire such as ”are you making more than 3000 euro per month?” with a binary answer should be easier to ask the patient rather than asking ”how much” which could be potentially sensitive to the person. Secondly, people with ethnicity Polish, Turkish, Antillean and Aruba, and Morokoans should also be on the radar of the general practitioners. Especially for Polish and Turkish which according to the modeling result shown to be the most vulnerable population among all the other.

Thus, by knowing this information, the author suggests the following to be done to services rationing, managing, and conserving :

1. Try to shift the focus of decision making for cardiovascular risk prevention in more specific groups of people rather than covering the whole population. People that have age higher than 50 and/or live below the poverty line, with the ethnic minority that mentioned above could be the focus in judging some medical treatment for cardiovascular diseases prevention.

2. Prioritizing the mentioned susceptible population by allowing them to have extra followed up visit if they reported some early symptom of cardiovascular disease. Two things can do this, follow up visit after a while, or general practitioners (or its delegation) could call them by phone to ask for their continuation of the symptoms.

3. Allowing longer consultation time for these susceptible population to let them describe more detailed about what they are experiencing, not only the latest symptoms but also other strange symptoms that they experienced for the past year that could potentially be related to cardiovascular disease.

4. Employing different queuing technique to prioritize the susceptible group first instead of ”first come first serve” based queuing technique when making an appointment with the general practitioners.

5. Providing different information and/or advice to each susceptible population, especially for different ethnicity. For example, a general practitioner can choose to disclose information regarding the higher cardiovascular risk in Polish ethnic group to Polish patient while general practitioner can also choose to not reveal information to Dutch patient about the smaller risk among people with Dutch ethnic group. In addition to that, by gathering more information about the behavior of these specific susceptible group of people, especially their lifestyle and dietary habits, a general practitioner can also give more tailored advice to the patients.

6. Since there is less number of people that need to be prioritized for cardiovascular risk prevention, general practitioners can try to allocate their "slack time" to respond to an emergency or unpredictable situation that susceptible population may have.

7. Delegating some of the work related to cardiovascular diseases prevention to other public servants that have a higher or lower level. The levels refer to either expertise level of hierarchical level. For example, general practitioners could delegate some monitoring tasks to the nurses while they can also refer the patients to the medical specialist for a more specific case. On the other hand, in terms of hierarchical, general practitioners could also try to collaborate with a municipality to provide a cardiovascular diseases prevention program for all of the susceptible groups mentioned above since general practitioners do not have the human and financial resources to do this alone. There is at least two policy that can be the foundation of collaboration between general practitioners and municipalities, Public Health Act (*Wet publieke gezondheid*, Wpg) and Long-term Care Act (*Wet Langdurige Zorg*, Wlz).

Cooperating and regulating patients to follow the standard procedure

Since the street-level bureaucrats interact directly with the clients, it is common that they need to make standard procedure on how to handle clients interaction. The same applies to general practitioners, that used multiple guidelines that strictly regulated by NHG. In NHG guidelines for cardiovascular diseases prevention, the information needed for general practitioners to make the decision for what medical treatment or recommendations are needed to be given to the patients. Moreover, it is without a doubt that medical measurements are the primary indicator that highlighted in the NHG guidelines for identifying the potential risk of cardiovascular disease among individual. However, given the fact that there is a low number of medical measurement report on the ELAN dataset, shows how scarce the information that general practitioners have when performing cardiovascular diseases prevention act. Commonly, there is three main reasons for data scarcity in medical measurement. Firstly, there is not enough evidence for general practitioners to refer patients to get a medical check-up. Secondly, some patients are recommended by general practitioners to take the medical measurement, but does not shows up in the medical examination. Thirdly, there is an individual who never visits a general practitioner and never requests a medical check-up even though it is recommended to take a medical check-up at least once in a year.

Hence, to improve the compliance of the patients to have a medical check-up, general practitioners could try to use the result from this research to encourage them to get medical measurements. This is potentially useful in the second points mentioned above. In particular, by imposing psychological pressure on susceptible patients by giving them information about their high risk compared with other groups. Thus, hopefully, the susceptible patients will start looking after themselves and show to a medical check-up more often. However, it is still tricky for a general practitioner to communicate the risk to a patient that never visit. Then, the common ways to approach this behavior is to give financial compensation (free charge) for medical measurements or to make the measurement compulsory before the visits. Although, this option requires general practitioners to collaborate with higher levels organizations such as municipality since that is where the potential funds that could help them.

Managing the consequences of routine practice

It is not that uncommon that the street-level bureaucrats day to day operation that based on the standard procedure does not work or even generate unpredictable reactions from clients. If such a case happens, they will have to use other practices or even act based on their own belief and values from their own experience. This research, in particular, shows the reflection of how the general practitioners deal with the consequences of the standard practice that fails to conform to their beliefs. This is because this research request came from the general practitioners association in the Hague. By slowly gathering the piece of evidence that confirms their beliefs, it has become easier for general practitioners to perform their routine without the need to worry about giving information to the patients solely by their ”hunch”. However, communicating the risk based on the socioeconomic status also potentially add another complication to the standard practice. First, identifying the socioeconomic status of the individual (by asking them) might make the individuals feel uncomfortable around the general practitioners, thus making them avoid the medical practice altogether. Hence, on the many solutions that general practitioners could do is try to get collaboration with the ”data broker” such as Central Bureau of Statistics (CBS), so that the information for the individual before the visit could be checked. Secondly, a patient could also feel threatened, attacked, or hostile when the treatment is separated by how much money they make and/or especially if it is separated by ethnicity. They may feel it is unfair for them and thus act hostile towards general practitioners, thus making it even harder for them to perform their routine later on. Thus, proper additional research, rules, and policy need to be in place before general practitioners used socioeconomic status in their real-world practice.

6.2.1. Policy recommendations

Summarizing all points that mentioned above, there is one policy advice that can be drawn from this research according to the issues of what the general practitioners as decision-makers to cardiovascular diseases prevention act, these are:

”*Focus of cardiovascular diseases prevention program should be in the group of people with age more than 50 years old that live below the poverty line and are part of ethnic minority group especially, Turkish, Moroccans, Antillean and Aruba, and Polish people*”

6.3. Discussion

In this section, the author opinion and thoughts about the whole process research and modeling process will be discussed. Starting from research fundamentals, which then continues to discuss several points that the author observed in each modeling steps according to the CRISP-DM framework. Finally, the model limitation and future research will then be discussed in the last section to conclude the reflection of the research.

6.3.1. Research fundamentals

Fundamentally, this research used two main concepts to structure the whole research, street-level bureaucrats theory and CRISP-DM framework. First, by using street-level bureaucrats theory, it is much easier during the modeling process to reflect on the practicalities of the variables and the modeling results interpretation since the author can link and compare directly to the ”nature” and ”behavior” of the street-level bureaucrats. In addition to that, it is also much easier to construct the recommendations and policy advice that is tailored to the general practitioners by using this theory. Thus, the main benefit to using the street-level bureaucrat's theory comes from the fact that we can learn some general pattern and behavior that public servants have related to the service that they provide and how they perform it. Secondly, CRISP-DM framework that is used in this research is useful to employ since it is a quite elaborate yet straightforward framework that shows the transparency of the whole modeling process. However, the author thinks that there should be some adjustment that needs to be done for the last step in the CRISP-DM framework for academic research. Since the deployment phase in most research is probably not within the scope of the research. For example, if the research purpose is to explore the several different modeling practices without any practical goals in mind, it is merely not feasible to do deployments phase since no product needs to be deployed.

6.3.2. Understanding the system of interest

The first modeling process begins by gathering as much information as possible related to the system of interest. In this research, by gathering knowledge about how the healthcare systems work and especially the scope of how general practitioners are regulated the author gain insight in how the data is collected and some likely explanation about the variable values. For example, in variable INHPOPIIV there is a categorical variable of institution household which we know from the Chapter 2 is regulated through Long term act which is funded by the national government and governed by municipalities. Thus knowing this fact, it should be expected that people who belong to this type of household, regardless whether they have observed income or not, should not have that many different prognostics since eventually they are treated the same. This is proven in result with INHPOPIIV 3 ad 8 being significant and have the same prognostics. Although the result suffers from overfitting and the interpretation might not be correct, but the author is guessing that both of them will have the same interpretation in the end. In addition to that, having knowledge of the healthcare system provides to be useful, especially when handling some outliers even understanding about how possible missing values are produced, especially in the ELAN dataset. One can infer that given several steps that a patient need to go through for just getting some simple medication, the patient might lose motivation to get a necessary medical check-up thus explained the missing values in the dataset.

Subsequently, having information about how the general practitioners work (Checkland, 2004; Lugtenberg et al., 2009; Zha.nl, 2019) provide insight into this research, especially during variables selection in the modeling process. Some simple, straight forward, and self-explanatory variables are definitely one of the main criteria when eliminating a list of variables during the modeling process. Therefore, most common tools such as principal component analysis (PCA) (Harrell Jr, 2015) for dimensionality reduction is not encouraged to use in this research since the interpretability of the variables is going to be lost when it is merged into several confounding factors. In the end, the goal of this research is for communicating some relevance socioeconomic status variables that can identify the susceptible population to exposure of first cardiovascular diseases. If the variable is hard to interpret, or even hard

to get the information from the patient, then there is no point in adding that variables to the model.

Lastly, understanding the system of interest is particularly useful, especially when modeling Discrete Bayesian Network. It is well known by multiple works of literature and medical expert on how medication and basic demographics such as sex and age-connected through the medical measurements, but there is not that much literature that stated how the socioeconomic status dependency to each other and to various medical measurements. As it is demonstrated by Scutari and Denis (2014) and Bandyopadhyay et al. (2015), it is not that uncommon in Discrete Bayesian Network to use ”knowledge expertise” to define the model structure, hence the name ”handcrafted model” usually associated with this type of modeling practice.

6.3.3. Modelling process

Usually, when survival analysis in the field of medical research is conducted, the data are collected based on clear study design and objective. The two most important thing to consider is the follow-up time (Clark et al., 2003) and the sample size for each controlled group population (Bradburn et al., 2003a). Failed to consider this during the research study design, could lead to several biases in the survival analysis, thus reducing the credibility of the model. Therefore, commonly in medical research, the study design comes first; then, the data are collected afterward. However, that is not the case in this research. The study design and purpose are actually defined by the data that arbitrary collected in various general practitioners and statistical institutions which might not even fit for doing survival analysis. Therefore, there is a lot of modeling ”tricks” that employed in this research to make the model that is chosen to fit with the dataset that used.

6.3.4. Model evaluation and interpretation

First and foremost, evaluating the semi-parametric Cox proportional hazards model is definitely such a labor-intensive work due to the fact that each covariate needs to have proportional hazards ratio across time. Although in the model validation results, only two iterations are shown. In actuality, the author did more than ten iterations with many different combinations of techniques to be able to somehow satisfied the proportional hazards assumptions in the model. Even after the proportional hazards assumptions are satisfied, the interaction term that is used as a trick to satisfied the Cox model assumptions make the modeling result interpretation harder than without interaction since now there is more dimension to cover in the interpretation. However, for Discrete Bayesian Network, the result are much more clear and easier to understand although some additional check is required for how the model structure is defined.

Frequentist vs Bayesian approach In this paragraph, the author wants to reflect on the modeling results that are produced by Discrete Bayesian Network in the case of ethnicity hazards ratio. Although it is not clearly stated in the result or conclusion, the result from the Bayesian approach are actually confirm several general practitioners hunch about the perceived higher risk in people with Polish ethnicity background. This is due to several conversations that the author had with general practitioners, in which they repeatedly mentioned that GP sees more Polish people getting cardiovascular diseases event much younger compared to the other ethnicity. However, not so many literature explore this issues in the Netherlands and if there is a research conducted for cardiovascular diseases in the ethnic minority groups, they most often primarily focus more on other ethnic minority groups simply because they have bigger sample size (Perini et al., 2018).

What author find most surprising is that how Polish ethnic group in the Discrete Bayesian Network result with maximum likelihood estimation are considered as the people with the lowest hazards ratio, jump up into the category of the population with the highest risk of cardiovascular diseases. Even though the imaginary sample size (iss) that is used in calculating the parameter for the Discrete Bayesian Network is relatively small (10) compared to the number of rows (377,885). This proves how bias frequentist approach could be in dealings with unequal sample size as opposed to the Bayesian approach.

6.3.5. Research limitations and recommendations

The main limitation of the whole modeling process in this research mainly comes from two reasons, the dataset that used and the CBS environment that this research bound to. The combined dataset from ELAN and CBS dataset proven to be insufficient for the research purposes due to a large number of missing values in the dataset. In addition to that, the CBS environment in which the author needs to work with the data and models always restrict some of the modeling processes that sometimes limits the potential use of some of the modeling techniques.

Study design and preprocessing

1. To ensures the credibility of the model, the author thinks that at least the dataset should not be missing for more than 10 percent, especially for the medical measurements. As can be seen in this research, there is no medical measurement that adds significance to the first cardiovascular event as opposed to the socioeconomic status which almost half of the variables add significance to the dependent variable. However, the author believes that if the medical measurement value has the same data sparsity status as socioeconomic status, the significance of the medical measurement variable will rise respectively to the dependent variable.

2. The events of interest that are chosen for analysis are the first cardiovascular diseases. However, it is possible that some older people already have experienced a cardiovascular event before the study start (2011) and therefore, the definition of ”first” could potentially not hold anymore. Therefore, it is probably useful in the future to check the past cardiovascular events; therefore, that particular individual can be excluded from the dataset. One of the ways to identify whether the patient already had cardiovascular events or not, could be through medication code that they are taking. According to an expert, there is some specific medication that a patient can and should take after having a cardiovascular event. By taking this medication code, the researcher can then exclude patients from the dataset since they already experience an event before. However, research also needs to be careful when excluding some patients as they could fall into the truncation issues if the number of cases for exclusion is relatively high (Lamarca et al., 1998).

3. The medication label that is used in this research appears to be too broad since it makes no separation between (anti)hypertensive medication which intended to lower blood pressure and lipid-lowering drugs that used to lower cholesterol (LDL). Therefore it is recommended in the next research to make a separation for two types of medication that is related to cardiovascular diseases. In addition to that, the medication count that is used in this research is not really accurate in the repeated measurement since the values for each repeated measurements are an aggregation of the yearly sum of cardiovascular medication prescription.

4. During the data preprocessing, always check the data sparsity and duplicated values before performing some analysis. This research deals with duplicated values by taking the values that have the least number of missing values count. But probably there is a better way out there to deal with multiple duplicated values that account for also variable of interests when removing the duplicates. In addition to that, it is also recommended to a used enhanced data frame with optimizing memory usage and garbage collection (in R it is called datatable) since it is way faster in processing the syntax compared to regular dataframe or other data type. And probably try to separate the dataset into smaller meaningful batches; therefore, some function or even the model can be run in a much more faster way.

Recommendations for study design Summarizing all points mentioned above, these are the author recommendation for designing cardiovascular disease risk prediction model study in the future :

- Consider for each covariate (especially the categorical) have considerably sufficient sample size (more than or equal than 10 percent of the largest group of the population) with at least ten events
- Try to exclude people below 40 years old before the study is conducted. However, if there are a high number of people that have age below 40 years old, then the model should take into account the bias that comes from truncation.
- Consider to do open cohort as opposed to close cohort. This research, in particular, used a closed cohort due to the dataset that is used. However, if the real study is going to be conducted, it is possible to keep track of people that goes in and out of the study.
- Then, it is also recommended to collect lifetime data, not just merely for some period of time. This should be possible if there is a possibility to run the national level electronic health record programme.
- Strict follow up for the measurement is definitely recommended to ensure data quality, thus increasing the credibility of the model. This is to minimize the use of imputed dataset since the significance and variability of the prediction model is highly dependent on the actual data, not simulated ones.
- If it is possible, try to minimize the number of people that drop out during the study is conducted to minimize the number of people that are censored.

Modelling

1. In multivariate analysis of the Cox proportional hazards model, almost all of the percentile group are treated as a continuous variable. This means that the author assumes that there is linearity for different levels of the group. And this is not always the case, as can be seen in the result of prosperity variables. Hence, two recommendations are made if regression analysis is going to be performed for this type of ordinal categorical variables. The first recommendation is to perform the linearity test while second suggestions refer to treat the variable as a factor variable and then make a dummy variable out of each level in these factor variables. There are several linearity tests that can be inspected by either in the form of statistical test or graphical representation (Harrell Jr, 2015).

2. As opposed to the first point, using a dummy variable to model categorical covariates makes the interpretation of the regression in the Cox proportional hazards model sometimes contradictory since it used one level as the reference for the rest of the categorical dummy variables. For example, in the interpretation of INHPOPIIV 2 (assuming that it is not overfitting), the higher value of INHPOPIIV 2 related to good prognostic which means that INHPOPIIV 1 related to worse prognostics compared to INHPOPIIV 2. On the other hand, INHPOPIIV 3 related to bad prognostic, which means that INHPOPIIV 1 related to better prognostics compared to INHPOPIIV 3. So given these two interpretation of INHPOPIIV 1, it is hard to grasp the nature of INHPOPIIV 1 since it appears as the reference for the other dummy variable as well. Therefore, it is recommended when dealing with categorical variables, reducing the number of levels can provide to have much more meaningful interpretation compared to categorical variables with a high number of levels.

3. By definition, there is a high chance that all of the socioeconomic statuses that used in this research are redundant and correlate with each other. Unfortunately, due to the limitation in computational power, this research could not run the multicollinearity test due to the high number of variables used. Therefore, it is recommended for future research to perform multicollinearity test to check the redundancy of some variables as part of model validation before interpreting the result.

4. Bradburn et al. (2003a) suggest that there is sample size consideration for each covariate. It is stated in their paper that ”the power of survival analysis is related to the

number of events rather than the number of participants". And it is recommended that at least 10 number of events should be available in each covariate. This assumption is particularly important to consider especially for a categorical variable that split different levels into a dummy variable. As can be seen in this result interpretation, there are a lot of variables that are over and underfitting, especially for categorical variables. Therefore it is suggested that simple frequency count to check the number of events for each different group of levels are necessary to be performed. If the levels or covariates contains less than 10 number of events, those covariates can either be combined with other covariates, or it can simply be excluded in the regression analysis.

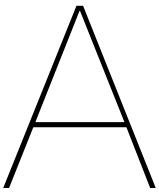
5. During the modeling process of parametric models in this research, Gompertz distribution appears to be the best fit with the dataset that is used. However, due to computational resources and R package limitation, currently, it is impossible for this research to use Gompertz distribution to model the cardiovascular risk for CBS and ELAN dataset. Hence, the author thinks that for further research, Gompertz parametric model for survival analysis could be used in the future research given that the time axis that is chosen is age and the event of interest is the first cardiovascular event.

6. It also recommended by literature (Bradburn et al., 2003b; Wang et al., 2019) that in the case that the Cox proportional hazards modeling assumptions do not hold, maybe it is better to use the Accelerated Failure Time (AFT) parametric model. AFT is the type of parametric model which the parameter of the distribution can replicate the various type of distribution such as the Log-Normal, Log-Logistic, Generalised Gamma, and Weibull. In the future AFT, might a good comparison as opposed to Gompertz or Cox model if both of the models are deemed to be inappropriate for the research.

7. Lastly, in Discrete Bayesian Network, the assumptions that each repeated measurement are considered to be different individuals seem to underestimate the hazards ratio of some of the variables. It is encouraged for future research, to try to use different assumptions for Discrete Bayesian Network, in particular taking the suggestion by Bandyopadhyay et al. (2015) which either just take people that have an event, or take one or two individual that censored and have an event. It is also recommended to add temporal assumptions into the model if the data have at least 80 percent of observed values since the dependencies of each variable across time is logical to model when the data is not imputed. The author suggests that to be able to perform this, the temporal assumption can be aggregated to yearly repeated measurement to discretize the time as opposed to the current data which recorded based on the event-based time discretization.

Recommendations for modelling Thus, for the modeling approach, few recommendations can be summarized from all the points mentioned above. These are :

- There will inevitably be censored individual, and it is useful to perform survival analysis for a censored individual to define the weight of the censored relatively to people who experience the event
- Put different assumption for Discrete Bayesian Network. Especially in the allocation of priors that and imaginary sample size (iss) in Bayesian approach for parameter learning. The value of iss can be matched up with the weight from the survival analysis for censored data.
- Perform a competing risk survival analysis. Which means that censored individual is not considered as risk-free individuals (not merely censored), but the censored individual still stays in the cohort, and there are multiple events occurred in the study.



Appendix

A.1. Data dictionary

This appendix section contains the data dictionary and variable explanation of each variable that is used. The first subsection list all of socioeconomic variables in CBS microdataset that could potentially meaningful or used as a variable in the the modelling part. The second subsection contains all variable that already consulted and selected that could be used for the modelling purposes.

A.1.1. CBS microdata dictionary

Table A.1: Portion of CBS data dictionary part 1

Variable	Filename	Variable name	Variable values	Description
Ethnicity	GBANATIONALITEITBUS Key: RINPERSOON / STRING	GBANATIONALITEIT1	[XXXX] / STRING	primary nationality (legal citizenship)
		GBANATIONALITEIT2	[XXXX] / STRING	secondary nationality (legal citizenship)
		GBATYPENATIONALITEIT	[-,1,2,3,4, 5, 6, 7, 9, .] / STRING	the number and combination of nationalities of a person
	GBAPERSOONTAB Key: RINPERSOON(S) / STRING	GBAGEBOORTELAND	[XXXXX] / STRING	Country of birth of individual
		GBAGEBOORTELANDMOEDER	[XXXXX] / STRING	Country of birth of individual mother
		GBAGEBOORTELANDVADER	[XXXXX] / STRING	Country of birth of individual father
		GBAHERKOMSTGROEPERING	[XXXXX] / STRING	Migration Background (CBS definition): Country with which a person is connected based on the birth of the parents or themselves
		GBAGENERATIE	[-,0,1,2] / STRING	Generation (first and second generation immigrant background)
Income	INHATAB Key: RINPERSOONHKW / STRING	INHARMEUR	[-2, -1, 1, 2,….,999] / NUMERIC	Income with respect of the European poverty line during the year
		INHARMEURL	[-4, -2, -1, 1, 2,….,999] / NUMERIC	Income to the European poverty line in the last four years
		INHARMLAG	[-3, -2, -1, 1, 2,….,999] / NUMERIC	Income with respect to the low-income in the past year
		INHARMLAGL	[-4, -3, -2, -1, 1, 2,….,999] / NUMERIC	Income with respect to the low-income in the past four years
		INHARMSOC	[-3, -2, -1, 1, 2,….,999] / NUMERIC	Income with respect to the minimum policy in the year
		INHARMSOCL	[-4, -3, -2, -1, 1, 2,….,999] / NUMERIC	Income relative to the policy minimum in the last four years
		INHBBIHJ	[11,12,13, 14,….99] / STRING	Main source of income of the household
		INHEHALGR	[1,2,3,8,9] / STRING	Ownership private households on 1 January of the year
		INHP100HBEST	[-2,-1,1,2 ,….,100] / NUMERIC	Private households divided into 100 groups of equal size based on disposable income
		INHBESTINKH	[XXXXX] / NUMERIC	Disposable household income
		INHGESTINKH	[XXXXX] / NUMERIC	Standardized disposable household income
		INHP100HBESTES	[-3,-2,-1, 1,2….,100] / NUMERIC	Private households divided excluding student households in 100 groups of equal size based on disposable income
		INHP100HBRUT	[-2,-1,1,2 ,….,99,100] / NUMERIC	Private households divided into 100 groups of equal size based on gross income
		INHP100HGEST	[-2,-1,1,2,…., 99,100] / NUMERIC	Private households divided into 100 groups of equal size based on the standardized income
		INHP100HGESTES	[-2,-1,1,2,…., 99,100] / NUMERIC	Private households divided excluding student households in 100 groups of equal size based on the standardized income
		INHP100HPRIM	[-2,-1,1,2,…., 99,100] / NUMERIC	Private households with primary income divided into 100 groups of equal size on the basis of the primary income

Table A.2: Portion of CBS data dictionary part 2

Variable	Filename	Variable name	Variable values	Description
Income	INPATAB Key: RINPERSOONHKW / STRING	INPAZLF4240P	[XXXXX] / NUMERIC	The amount that can bring an entrepreneur deducted from profits
		INPEMEZ	[0,1,8,9] / STRING	Classification Variable (socio-economic) situation where the personal net income from employment or self-employment is higher than the net income support for a single
		INPEMFO	[0,1,8,9] / STRING	Classification Variable (socio-economic) situation where the personal net income from employment or their own company as well as from social insurance is higher than the low-income threshold for a single person
		INPIMPINK	[10,20,30, 32,...99] / STRING	The source of data on income are made
		INPIMPZELF	[10,20,30, 32,...99] / STRING	The source from which the data are composed of independent entrepreneurs
		INPP100PBRUT	[-3,-2,-1 ,…..,100] / NUMERIC	Gross personal income divided into 100 groups of equal size of individuals with income in private households
		INPP100PPERS	[-3,-2,-1 ,…..,100] / NUMERIC	Personal income divided into 100 groups of equal size of individuals with income in private households
		INPPERSBRUT	[XXXXX] / NUMERIC	Gross personal income
		INPPERSINK	[XXXXX] / NUMERIC	The personal income includes the following items of gross income of a person: earned income, income from own company pension income insurance and social security benefits (excluding family allowances and child budget). Income insurance premiums have been deducted
		INPPERSPRIM	[XXXXX] / NUMERIC	The primary personal income includes gross income of a person from employment or from private enterprise. Labor income consists of gross wages (including employee and employer's contributions for social insurance), bonus and reward work that is not performed in employment. Also, wages in kind such as the value of the private use of the car of the employer is included in this. Income from own company is the reward of self-employed for the use of their labor and business assets
		INPSECJ	[11,12,13, 14,…..99] / STRING	Classification of a person by socio-economic category based on income sources in a year
		INPT1000WER	[XXXXX] / NUMERIC	Wage, salary, bonus, savings of a worker in the private sector. Also includes salary received from abroad. The amount includes the employee, but excluding employer's contributions for social insurance
		INPT1020AMB	[XXXXX] / NUMERIC	Wage, salary, bonus, savings of an official. The amount includes the employee, but excluding employer's contributions for social insurance
		INPTYPZLF	[0,1,2,3,4,9] / STRING	A person who works for his own account or risk their own business or practice. Notes to the definition Typing entrepreneur to the presence or absence of personnel and economic activity (production and services). These are people with an income tax return income tax return
		INPV3900INK	[XXXXX] / NUMERIC	Amount of income tax. The income tax as the tax payable on the income of the year. The amount is the balance of the (gross) income (IB) and the IB part of the tax credit (from 2001). If there is no assessment of income tax is imposed, the income tax is equal to the pre-charges in the form of income tax and dividend
Occupation	BAANKENMERKENBUS Key: RINPERSOON / STRING	SOORTBAANID	[1,2,3,4,5,9] / STRING	Job type (Position)
		ARBEIDSRELATIE-BAANID	[1,2] / STRING	Employment status
		CAOSECTORBAANID	[1000, 2000, 3000, ….. , 3800] / STRING	CAO Sector Code
		SECTBAANID	[00,01,02, ….,99] / STRING	Job sector
		AUTOVANDEZAAK-BAANID	[0,1] /STRING	Company Car

Table A.3: Portion of CBS data dictionary part 3

Variable	Filename	Variable name	Variable values	Description
Education	HOOGSTEOPLTAB Key: RINPERSOON / STRING	OPLNRHB	[0000001, ..., 9999999] / STRING	Education Code highest grade education
		OPLNRHG	[0000001, ..., 9999999] / STRING	Education highest number of training
		OPLNIVSOI2016A - GG4HBMETNIRWO	[—,1110, 1111,1112, ….., 9999] / STRING	Education in 18 categories highest grade education
		OPLNIVSOI2016A -GG4HGMETNIRWO	[—,1110, 1111,1112, ….., 9999] / STRING	Education in 18 categories highest qualifications
		RICHTdetailISCEDF-2013HBmetNIRWO	[0000, 0010, 0011, 0020, ..., 9998] / STRING	Education Direction highest achieved education incl. Estimations for Not in Education Registers observed
		RICHTdetailISCEDF-2013HGmetNIRWO	[0000, 0010, 0011, 0020, …., 9998] / STRING	Education Direction highly tracked education incl. Estimations for Not in Education Registers observed
		BRONOPLARCHIEFHB	[01, 02, 03, 04, 05, 06, 07, 08, 09, 10] / STRING	Abbreviation source highest achieved education
		BRONOPLARCHIEFHG	[01, 02, 03, 04, 05, 06, 07, 08, 09, 10] / STRING	Abbreviation source highly track ededucation
Address	GBAADRESOBJECTBUS Key: RINPERSOON / STRING	GBADATUMAANVAN-GADRESHOUDING	[XXXXXXXX] / STRING	Date on which a person is registered in the municipal
		SOORTOBJECT-NUMMER	[B,H,D,O] / STRING	The code gives of Residence ID that is a unique identifier
	VSLGWBTAB Key: RINPERSOON / STRING	GemJJJJ	[XXXXXX] / STRING	Municipal Code on year
		WCJJJJ	[XXXXXXXX] / STRING	District Code on year
		BCJJJJ	[XXXXXXXXX] / STRING	Area Code of that year
	VSLPOSTCODEBUS Key: RINPERSOON / STRING	RINOBJECT-NUMMER	[XXXXXXXX] / STRING	This number identifies a recipient residing
Facilities access	NABIJHEIDHORECATAB Key: RINPERSOON / STRING	VZAFSTANDCAFE	[XXXXXX] / NUMERIC	Distance in meters from an address to the nearest cafe, calculated on the road
		VZAANTCAFE05KM	[XXXXXX] / NUMERIC	Number of cafes within a radius of 5 km
		VZAANTCAFETARIA05KM	[XXXXXX] / NUMERIC	The number of cafeterias that from the address within 5 km of the road is accessible
		VZAANTRESTAU05KM	[XXXXXX] / NUMERIC	Number of restaurants within 5 km of the road to reach
Martial status / household	INHATAB Key: RINPERSOONHKW / STRING	INHAHL	[1,2,3.....,99] / NUMERIC	Number of persons in the household
		INHAHLMI	[1,2,3.....,99] / NUMERIC	Number of household members with personal income
		INHPOPIIV	[1,2,3.....,9] / STRING	Household type to indicate whether a household has an income observed
		INHPRIMINKH	[XXXXXX] / NUMERIC	The primary income includes income from employment, income from own business and income
		INHSAMAOW	[11,12,13,….,88] / STRING	Household composition, membership in pension age
		INHSAMHH	[11,12,13,22,…,88] / STRING	Characterization of a household based on the relationships of the individuals within a household and sex and age
		INHUAFTYP	[0,1,2,3,4, 8,9] / STRING	Main benefit of private households
	INPATAB Key: RINPERSOONHKW / STRING	INPPOSHHK	[1,2,3,4,5,6, 7,9] / STRING	The position of a person in a household to the main income earner of the household
	VEHTAB Key: RINPERSOONHKW / STRING	VEHP100WELVAART	[-2,-1,….., 100] / STRING	Private households divided into 100 groups of equal size based on the power and the standardized income. Institutional households and households whose income is unknown, are not included in the percentiel (no target population)
		VEHP100HVERM	[-2,-1, 1,2, 3,…..,100] / NUMERIC	Private households divided into 100 groups of equal size on the basis of the power
		VEHW1120ONRH	[XXXXXX] / NUMERIC	The total value of the home and other property of a household
		VEHW1121WONH	[XXXXXX] / NUMERIC	Value of the property owned by a household and used as a main residence
		VEHW1122OGOH	[XXXXXX] / NUMERIC	Total value of property of a household, excluding property that serves as a main residence

A.1.2. CBS and ELAN data dictionary

Table A.4: Biological characteristics data dictionary

Biological characteristics

Variable	Filename	Column name	Column values	Description
Age available from 1990	GBAPERSOONTAB Key: RINPERSOON / STRING	GBAGEBOORTEJAAR	[XXXX] / STRING	Birth year
		GBAGEBOORTEMAAND	[XX] / STRING	Birth month
		GBAGEBOORTEDAG	[XX] / STRING	Birthday
Sex available from 1990		GBAGESLACHT	[-,1,2] / STRING	Sex of a person
Body mass index (BMI) available from 1990	MEETWAARDENCBKV1 (ELAN dataset) Key: gp_patidf_crypt / STRING	gp_examne - QUETAO	[XXXX] / STRING	Quetelet index (BMI) patient in kg/m^2

Table A.5: Measurements and medical records data dictionary

Measurement and medical records

Variable	Filename	Column name	Column values	Description
Smoking status available from 2009	MEETWAARDENCBKV1 (ELAN dataset) Key: gp_patidf_crypt / STRING	gp_examne - ADMIAQ, SIPDAQ, PPPDAQ, SGPDAQ, SRDAAQ, ROVWAZ, MOSRAQ, RESRAQ, ROOKAQ, STOPAQ	gp_exaval1 and gp_exatxt1 - [multiple string description - yes, no, never, unknown, etc] / STRING	Multiple definition of smoking
Total cholesterol available from 2009		gp_examne - CHOLB	gp_exaval1 - [XXXX] / STRING	Total cholesterol in mmol/l
High-density lipoprotein cholesterol (HDL-C) available from 2009		gp_examne - HDLB	gp_exaval 1- [XXXX] / STRING	HDL cholesterol in mmol/l
Low-density lipoprotein cholesterol (LDL-C) available from 2009		gp_examne - LDLB	gp_exaval1 - [XXXX] / STRING	LDL cholesterol in mmol/l
Triglycerides level available from 2009		gp_examne - TRIGB	gp_exaval1 - [XXXX] / STRING	Triglycerides (fat) in mmol/l
Systolic blood pressure (SBP) available from 2009		gp_examne - RRSYKA	gp_exaval 1- [XXXX] / STRING	Standard systolic blood pressure measurement in mmHg
Diastolic blood pressure (DBP) available from 2009		gp_examne - RRDIKA	gp_exaval1 - [XXXX] / STRING	Standard diastolic blood pressure measurement in mmHg
Diagnosis records available from 2009	EPISODECBKV1, JOURNAALCBKV1, MEDICATIECBKV1, VERRICHTINGENCBKV1 (ELAN dataset) Key: gp_patidf_crypt / STRING	icpc	[ICPC 1 code] / STRING	Classification of primary care diagnoses code
Death available from 2007	DO Key: RINPERSOON / STRING	OVLDAG, OVLMND, OVLYEAR	[XX], [XX], [XXXX] / NUMERIC	date of death
		PRIMOORZ	[ICD 10 code] / STRING	cause of death

Table A.6: Medication and medical treatment data dictionary

Medication

Variable	Filename	Column name	Column values	Description
(Anti)hypertensive treatment (Blood pressure related drugs) available from 2007	MEDICIJNTAB MEDICATIE (ELAN dataset) Key: RINPERSOON / STRING	ATC4, atc_code	ATC4 code] / STRING	Classification of drug groups based on four and six positions ATC code
Statins (Cholesterol related drugs) available from 2007				

Table A.7: Socioeconomic status data dictionary

Socioeconomic status

Variable	Filename	Column Name	Column values	Description
Ethnicity available from 1990	GBAPERSOONTAB Key: RINPERSOON / STRING	GBAHERKOMST-GROEPERING	[XXXXX] / STRING	Migration Background (CBS definition): Country with which a person is connected based on the birth of the parents or themselves
Income available from 2011	INHATAB Key: RINPERSOONHKW / STRING	INHARMLAG	[-3, -2, -1, 1, 2,….,999] / NUMERIC	Income with respect to the low-income in the past year
		INHARMSOC	[-3, -2, -1, 1, 2,….,999] / NUMERIC	Income with respect to the minimum policy in the year
		INHBBIHJ	[11,12,13,14,….99] / STRING	Main source of income of the household
		INHP100HBEST	[-2,-1,1,2,….,100] / NUMERIC	Private households divided into 100 groups of equal size based on disposable income
		INHBESTINKH	[XXXXX] / NUMERIC	Disposable household income
		INHGESTINKH	[XXXXX] / NUMERIC	Standardized disposable household income
		INHP100HBRUT	[-2,-1,1,2,….,99,100] / NUMERIC	Private households divided into 100 groups of equal size based on gross income
		INHP100HGEST	[-2,-1,1,2,….,99,100] / NUMERIC	Private households divided into 100 groups of equal size based on the standardized income
		INHP100HPRIM	[-2,-1,1,2,….,99,100] / NUMERIC	Private households with primary income divided into 100 groups of equal size on the basis of the primary income
	Own definition Key: RINPERSOONHKW / STRING	Armoedegrens	[0,1] / NUMERIC	Own definition of people live above and below poverty line based on INHARMLAG
	Own definition Key: RINPERSOONHKW / STRING	Gest Besteedbaar Inkomen	[1,2,3] / NUMERIC	Research own definition of people that have high, medium, and low income based on INHP100HGEST
	Own definition Key: RINPERSOONHKW / STRING	Welvaart	[1,2,3] / NUMERIC	Research own definition of people that have high, medium, and low prosperity based on VEHP100WELVAART
	Own definition Key: RINPERSOONHKW / STRING	Vermogen	[1,2,3] / NUMERIC	Research own definition of people that have high, medium, and low wealth based on VEHP100HVERM
Household status available from 2011	INHATAB Key: RINPERSOONHKW / STRING	INHAHL	[1,2,3.....,99] / NUMERIC	Number of persons in the household
		INHPOPIIV	[1,2,3.....,9] / STRING	Household type to indicate whether a household has an income observed
		INHSAMAOW	[11,12,13,….,88] / STRING	Household composition, membership in pension age
		INHSAMHH	[11,12,13,22,…,88] / STRING	Characterization of a household-based on the relationships of the individuals within a household and sex and age
	INPATAB Key: RINPERSOONHKW / STRING	INPPOSHHK	[1,2,3,4,5,6,7,9] / STRING	The position of a person in a household to the main income earner of the household
	VEHTAB Key: RINPERSOONHKW / STRING	VEHP100WELVAART	[-2,-1,…..,100] / STRING	Private households divided into 100 groups of equal size based on the power and standardized income. Institutional households and households whose income is unknown, are not included in the percentile (no target population)

A.2. Medical code label dictionary

Table A.8: ICD code

Death due to:	**ICD-10**
Hypertensive diseases	I10-I25
Angina Pectoris	I20
Acute myocardial infarction	I21
Subsequent myocardial infarction	I22
Complications after myocardial infarct	I23
Other ischemic cardiac disease	I24
Chronic ischemic cardiac disease	I25
Other heart disease	I46 - I51
Transient ischemic attack - TIA	G45
CVA's	I61, I62.1, I63 - I65, I67-69, except I67.1
Diseases arteries	I70 - I72

Table A.9: ICPC and ICD code

First episode of CVD	**ICPC-1**	**ICD-10**
Transient ischemic attack - TIA	K89	G45
Cerebrovascular accident (CVA)	K90	I64
Cerebral infarct	K90.03	I63
Acute myocardial infarct	K75	I21
Subsequent myocardial infarct	-	I22
Complications after myocardial infarct	-	I23
Other/chronic ischemic cardiac disease	K76	I24 and I25
Angina pectoris	K74	I20
Decompensatio cordis	K77	I50
Aneurysma aorta	K99.01	I71
Claudicatio intermittens	K92.01	I73.9

Subdiagnoses: K76.01, K76.02, K74.01, K74.02, K77.01, K77.02

Abreviations:

- ICPC -1 codes International Classification of Primary Care

- ICD-10 International Statistical Classification of Diseases and Related Health Problems

Table A.10: ATC code

Disease	Drug proxy (ATC code)
Cardiovascular disease history	At least one prescription of: antithrombotics (B01), cardiac glycosides (C01), antiarrhytmics, nitrates (C01), antihypertensives (C02), diuretics (C03), beta-blockers (C07), calcium antagonists (C08) or ACE inhibitors (C09)
Cardiovascular disease history	At least one prescription of: Antithrombotics (B01), cardiac glycosides, antiarrhytmics, nitrates (C01), antihypertensives (C02), diuretics (C03), beta-blockers (C07), calcium antagonists (C08), ACE inhibitors (C09) or low-dose aspirin
Cardiovascular disease history	At least one prescription of: Cardiac therapy (C01), antihypertensives (C02), diuretics (C03), beta-blocking agents (C07), calcium-channel blockers (C08), agents acting on the renin-angiotensin system (C09) or lipid-modifying agents (C10)
Previous atrial fibrillation or flutter	Prescriptions of digoxin or vitamin K antagonists
Cardiovascular disease history	Prescriptions of angiotensin-converting enzyme inhibitors, beta-blocking agents, low-dose aspirin, statins, calcium antagonists, other antihypertensives, diuretics or nitrates
Cardiovascular disease	Prescriptions of oral anticoagulants (B01AA03, B01AA04), cardiac glycosides (C01A), anti-arrhytmics (C01B), cardiac stimulants (C01C), vasodilators (C01D), beta-blocking agents (C07), calcium channel blockers (C08), angiotensin-converting enzyme inhibitors (C09A, C09B) or angiotensin II antagonists (C09C, C09D)
Cardiovascular disease	Prescriptions of central antihypertensives, beta-blockers, diuretics, calcium channel blockers, ACE inhibitors or angiotensin II receptor antagonists
Hypertension, cardial arrhythmia and ischemic heart disease	Prescription of diuretics (C03), beta-blockers (C07), calcium channel blockers (C08), angiotensin-converting enzyme inhibitors (C09), antiarrhytmics (C01A, C01B), or antianginal drugs (C01D)
Angina pectoris	At least two prescriptions of nitrates
Angina	At least two prescriptions of nitrates
Ischaemic heart disease	Prescription of nitrate (C01DA)
Established coronary disease	Use of nitrates
Presumed ischaemic heart disease	Coprescription of aspirin (B01AC06, N02BA01) and nitrate (C01DA)
Ischaemic heart disease	Three or more prescriptions of nitrate (C01DA) or nicorandil (C01DX16) with aspirin (B01AC06, N02BA01)
Heart failure	Three or more prescriptions of loop diuretics (C03C)
Hypertension	Three or more prescriptions of beta-adrenoceptor blockers (C07), diuretics (C03A, C03B) or calcium channel blockers (C07)
Heart failure	At least one prescription of loop diuretics (C03C)
Congestive heart failure	Two or more prescriptions of digoxin with diuretics
Atherothrombosis	At least one prescription of platelet aggregation inhibitors (B01AC)
Heart disease	Prescriptions of renin angiotensin systemic antagonists, beta and alpha blockers or cholesterol reducers and lipotropics
Ischemic heart disease	Various combinations of prescriptions of nitrate, aspirin, atenolol, statin and digoxin
Myocardial infarction	Prescription of nitrates, aspirin, statins or beta-blockers
History of cardiovascular event	Two or more prescriptions of vitamin K antagonists or thrombocyte aggregation inhibitors
Incident cardiovascular event	Two or more prescriptions of thrombocyte aggregation inhibitors

Source : Pouwels et al. (2016)

A.3. Cox model input

Table A.11: Socioeconomic variable summary

Variable name	Values	Definition
Armoedegrens	0 (reference)	People live below poverty line
	1	People live above poverty line
Gest Besteedbaar Inkomen	1 (reference)	People with low income
	2	People with medium income
	3	People with high income
INHARMLAG	0 - 1000	Income with respect to the low-income in the past year
INHARMSOC	0 - 1000	Income with respect to the minimum policy in the year
INHP100HBEST	0 - 100	Private households divided into 100 groups of equal size based on disposable income
VEHP100WELVAART	0 - 100	Private households divided into 100 groups of equal size based on the power and standardized income.
INHBBIHJ	11 (reference)	Salary
	12	Salary director and major shareholder
	13	Profit entrepreneur
	14	Income other independent
	21	Unemployment benefits
	22	Social assistance
	23	Social assistance benefits other
	24	Benefits illness / disability
	25	Pension
	26	Student grants
	30	Property income
INHPOPIIV	1 (reference)	Private household with income
	2	Private student household with incomes
	3	Institutional the household income observation unit
	7	Private household without income
	8	Institutional household, without income
	9	Privately, but not belonging to the household population
INHSAMAOW	11 (reference)	Single man from pension age
	12	Single woman from pension age
	13	Both married couple from pension age
	14	One married couple from pension age
	15	Other meerp. huish., at least one person over retirement age
	20	Household with income allowance, at least one person over retirement age
	31	Single, until retirement age
	32	Multi-person household, all individuals in retirement age
	40	Household with income allowance, all persons in retirement age
INHSAMHH	11 (reference)	Single person household, man to retirement age
	12	Single person household, man from pension age
	13	Single person household, woman to retirement age
	14	Single person household, woman from pension age
	21	Couple without children, main breadwinner to retirement age
	22	Couple without children, main breadwinner from pension age
	31	Couple with only minor children
	32	Couple with minor and adult children
	33	Couple with only adult children
	41	Single parent, only minor children
	42	Single parent, minor and adult children
	43	Single parent, only minor children
	51	Couple without children but with other resident(s)
	52	Couple with only minor children and other resident(s)
	53	Couple with minor and adult children and other resident(s)
	54	Couple with only adult children and other resident(s)
	55	Couple families, only minor child and others
	56	Couple families, minor and adult child and others
	57	Couple families, only adult child and others
	58	Other multi-person household
	71	Population in institutions, establishments and homes
INPPOSHHK	1 (reference)	Head of household without a partner
	2	Head of household with partner
	3	Married partner
	4	Unmarried partner
	5	Minor child
	6	Adult child
	7	Other household
Herkomst gehercodeerd	0 (reference)	Dutch
	1	Surinamese
	2	Turkish
	3	Marokoans
	4	Antillen and Aruba
	5	Indonesian
	6	Polish
	7	Other

Table A.12: Medical variable summary

Variable name	Values	Definition
GBAGESLACHT	1 2	Male Female
MEDICATION LABEL	0 1	Not prescribe cardiovascular medication Prescribe cardiovascular medication
MEDICATION COUNT	0 - 20	Number of medication prescribed
SMOKING STATUS	1 2	Yes No
QUETAO	5 - 35	Body mass index
RRSYKA	50 - 250	Systolic blood pressure
RRDIKA	30 - 160	Systolic blood pressure
CHOLB	0.3 - 8	Total Cholesterol
LDLB	1 - 3.8	Low density lipid
HDLB	0.3 - 1.2	High density lipid
TRIGB	1.3 - 5.8	Triglyceride

A.4. Bayesian model input

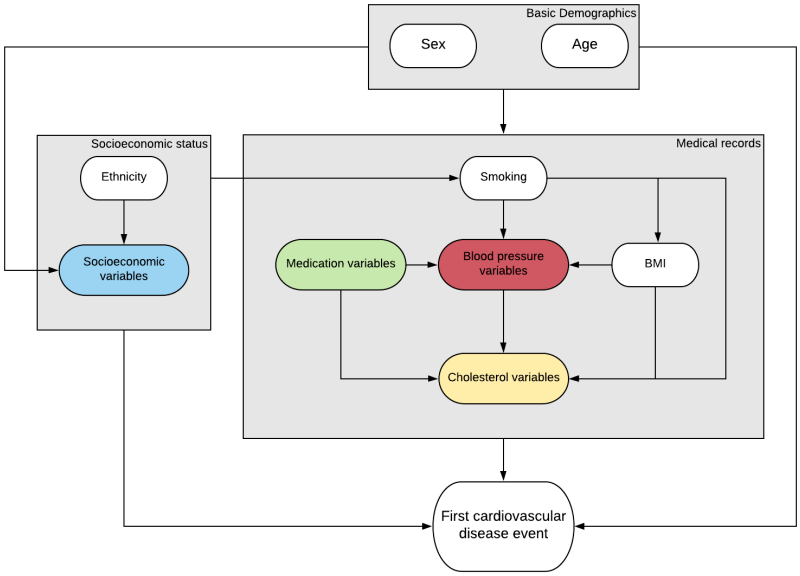


Figure A.1: Simplified version of Directed Acrylic Graph of Discrete Bayesian Model

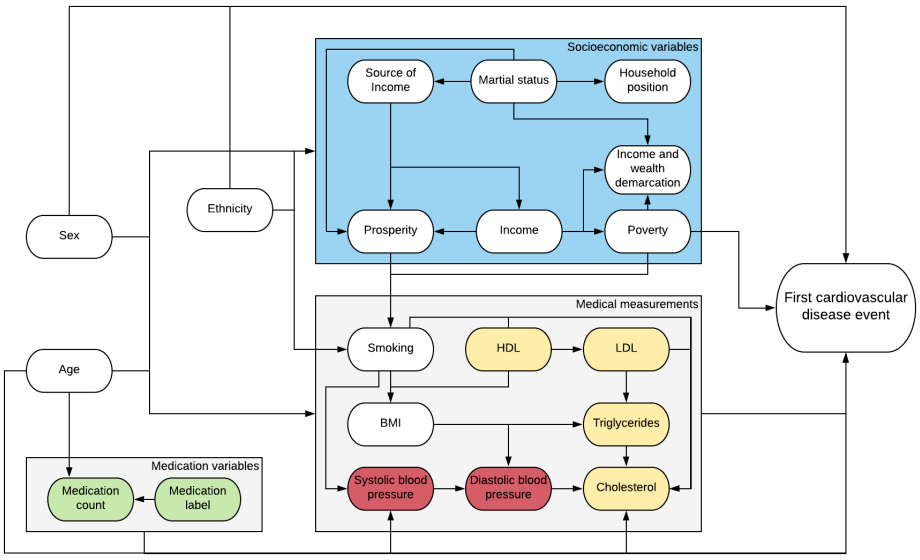


Figure A.2: Detailed version of Directed Acrylic Graph of Discrete Bayesian Model

Table A.13: Socioeconomic discretize variable summary

Variable name	Values	Definition
Armoedegrens	0	People live below poverty line
	1	People live above poverty line
Gest Besteedbaar Inkomen	1	People with low income
	2	People with medium income
	3	People with high income
Welvaart	1	People with low prosperity
	2	People with medium prosperity
	3	People with high prosperity
INHBBIHJ	11	Salary
	12	Salary director and major shareholder
	13	Profit entrepreneur
	14	Income other independent
	21	Unemployment benefits
	22	Social assistance
	23	Social assistance benefits other
	24	Benefits illness / disability
	25	Pension
	26	Student grants
	30	Property income
INHPOPIIV	1	Private household with income
	2	Private student household with incomes
	3	Institutional the household income observation unit
	7	Private household without income
	8	Institutional household, without income
	9	Privately, but not belonging to the household population
INHSAMHH	11	Single person household, man to retirement age
	12	Single person household, man from pension age
	13	Single person household, woman to retirement age
	14	Single person household, woman from pension age
	21	Couple without children, main breadwinner to retirement age
	22	Couple without children, main breadwinner from pension age
	31	Couple with only minor children
	32	Couple with minor and adult children
	33	Couple with only adult children
	41	Single parent, only minor children
	42	Single parent, minor and adult children
	43	Single parent, only minor children
	51	Couple without children but with other resident(s)
	52	Couple with only minor children and other resident(s)
	53	Couple with minor and adult children and other resident(s)
	54	Couple with only adult children and other resident(s)
	55	Couple families, only minor child and others
	56	Couple families, minor and adult child and others
	57	Couple families, only adult child and others
	58	Other multi-person household
	71	Population in institutions, establishments and homes
INPPOSHHK	1	Head of household without a partner
	2	Head of household with partner
	3	Married partner
	4	Unmarried partner
	5	Minor child
	6	Adult child
	7	Other household
Herkomst gehercodeerd	0	Dutch
	1	Surinamese
	2	Turkish
	3	Marokoans
	4	Antillen and Aruba
	5	Indonesian
	6	Polish
	7	Other

Table A.14: Medical discretize variable summary

Variable name	Values	Definition
CVD EVENT	Event Censored	First CVD events CVD events unknown
AGE	25-30 30-35 ..... 100+	Age between 25 - 30 Age between 30 - 35 ..... Age above 100
GBAGESLACHT	Male Female	Male Female
MEDICATION LABEL	Not treated Treated	Not prescribe cardiovascular medication Prescribe cardiovascular medication
MEDICATION COUNT	Never Small High	Number of medication prescribed 0 Number of medication prescribed < 5 Number of medication prescribed > 5
SMOKING STATUS	Yes No	Active smoker Non smoker
QUETAO	Ideal Pre-high Overweight Obese	5 < BMI < 20 20 < BMI < 25 25 < BMI < 30 above 30
RRSYKA	Low Normal High Critical	50 < SBP < 120 120 < SBP < 140 140 < SBP <160 above 160
RRDIKA	Low Normal High Critical	30 < SBP < 60 60 < SBP < 80 80 < SBP 100 above 100
CHOLB	Normal High Critical	0.3 < CHOL < 5 5 < CHOL < 8 above 8
LDLB	Low Normal Pre-high Critical	1 < LDL <1.8 1.8 < LDL < 2.5 2.5 < LDL < 3.5 above 3.5
HDLB	Low Good	0.3 < HDL < 1 above 1
TRIGB	Ideal Pre-high High Critical	below 1.7 1.7 < TRIG < 2.2 2.2 < TRIG < 5.6 above 5.6

B

Appendix

This appendix section contains the modelling results for Cox proportional hazards and Discrete Bayesian network. However, for full syntax of R code and several iteration table, please check the author github in https://github.com/AmmarFaiq.

B.1. Socioeconomic variables iteration

Table B.1: Cox proportional hazards model iteration summary for blacklisted (p-value > 0.25)

Analysis	Blacklisted variables (p-value)			
	Complete cases	Missing indicator	Single Imputation	Multiple imputation
Univariate	INHAHL (0.88)	Vermogen (0.28)	INHAHL (0.98)	INHAHL (0.94)
1st Multivariate	INHAHL (0.79) INHP100HPRIM (0.69) Vermogen (0.57) Welvaart (0.33) INHP100HBRUT (0.27)	INHAHL (0.91) INHP100HPRIM (0.82) Vermogen (0.59) INHBESTINKH (0.31) INHGESTINKH (0.37) Welvaart (0.31)	INHAHL (0.96) INHP100HPRIM (0.54) Vermogen (0.53) INHGESTINKH (0.30)	INHAHL (0.53) INHP100HPRIM (0.59) Vermogen (0.75) INHGESTINKH (0.29) Welvaart (0.35) INHP100HBRUT (0.27)
2nd Multivariate	Welvaart (0.42) INHP100HBRUT (0.30)	INHGESTINKH(0.41) Welvaart (0.39) INHBESTINKH (0.33) INHP100HBRUT (0.25)	INHGESTINKH (0.27)	INHGESTINKH(0.41) Welvaart (0.37) INHBESTINKH (0.33) INHP100HBRUT (0.31)
3rd Multivariate	INHBESTINKH (0.92)	INHBESTINKH (0.58)	INHBESTINKH (0.52)	INHBESTINKH (0.55)

Table B.2: Cox proportional hazards model iteration summary for whitelisted (p-value < 0.05)

Analysis	Whitelisted variables (p-values < 0.05)			
	Complete cases	Missing indicator	Single imputation	Multiple imputation
1st Mutivariate	Armoedegrens INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens INHARMSOC INHBBIHJ INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK Herkomst gehercodeerd	Armoedegrens INHARMLAG INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd
2nd Mutivariate	Armoedegrens INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens INHARMLAG INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd
3rd Mutivariate	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMLAG INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd
Final multivariate	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd	Armoedegrens Gest Besteedbaar Inkomen INHARMLAG INHARMSOC INHBBIHJ INHP100HBEST INHPOPIIV INHSAMAOW INHSAMHH INPPOSHHK VEHP100WELVAART Herkomst gehercodeerd

B.2. Cox proportional Hazards modelling result

Table B.3: Actual Cox Model (before adding interaction)

Variable Name	coef	exp(coef)	exp(-coef)	lower .95	upper .95	Pr(>|z|)	
MEDICATION COUNT	1.64E-02	1.02E+00	9.84E-01	1.01E+00	1.03E+00	0.000194	***
INHARMLAG	-9.45E-04	9.99E-01	1.00E+00	9.99E-01	1.00E+00	0.001912	**
factor(INHBBIHJ)12	-3.76E-02	9.63E-01	1.04E+00	6.89E-01	1.35E+00	0.826219	
factor(INHBBIHJ)13	-3.27E-02	9.68E-01	1.03E+00	7.82E-01	1.20E+00	0.763294	
factor(INHBBIHJ)14	6.42E-02	1.07E+00	9.38E-01	5.94E-01	1.91E+00	0.829864	
factor(INHBBIHJ)21	-1.44E-01	8.66E-01	1.16E+00	6.08E-01	1.24E+00	0.427029	
factor(INHBBIHJ)22	1.03E-02	1.01E+00	9.90E-01	8.15E-01	1.25E+00	0.924821	
factor(INHBBIHJ)23	-3.26E-01	7.22E-01	1.39E+00	4.94E-01	1.06E+00	0.093702	.
factor(INHBBIHJ)24	4.39E-01	1.55E+00	6.45E-01	1.30E+00	1.85E+00	7.15E-07	***
factor(INHBBIHJ)25	-1.01E-01	9.04E-01	1.11E+00	7.74E-01	1.06E+00	0.199185	
factor(INHBBIHJ)26	-1.20E+01	6.22E-06	1.61E+05	3.43E-06	1.13E-05	< 2E-16	***
factor(INHBBIHJ)30	-1.62E-01	8.51E-01	1.18E+00	6.49E-01	1.12E+00	0.24157	
INHP100HBEST	2.14E-03	1.00E+00	9.98E-01	9.98E-01	1.01E+00	0.369065	
factor(INHPOPIIV)2	-1.19E+01	6.50E-06	1.54E+05	3.88E-06	1.09E-05	< 2E-16	***
factor(INHPOPIIV)3	1.13E+01	7.90E+04	1.27E-05	1.78E+04	3.52E+05	< 2E-16	***
factor(INHPOPIIV)7	-5.79E-01	5.61E-01	1.78E+00	2.30E-01	1.37E+00	0.202758	
factor(INHPOPIIV)8	1.25E+01	2.74E+05	3.66E-06	2.36E+04	3.18E+06	< 2E-16	***
factor(INHSAMHH)12	-2.50E-01	7.79E-01	1.28E+00	6.19E-01	9.80E-01	0.033108	*
factor(INHSAMHH)13	3.18E-02	1.03E+00	9.69E-01	8.24E-01	1.29E+00	0.782389	
factor(INHSAMHH)14	-7.41E-02	9.29E-01	1.08E+00	7.32E-01	1.18E+00	0.540916	
factor(INHSAMHH)21	-1.81E-01	8.34E-01	1.20E+00	6.11E-01	1.14E+00	0.254912	
factor(INHSAMHH)22	-3.70E-01	6.91E-01	1.45E+00	4.93E-01	9.70E-01	0.032621	*
factor(INHSAMHH)31	-1.85E-01	8.31E-01	1.20E+00	5.73E-01	1.21E+00	0.328402	
factor(INHSAMHH)32	-5.28E-01	5.90E-01	1.70E+00	3.64E-01	9.55E-01	0.031856	*
factor(INHSAMHH)33	-1.25E-01	8.83E-01	1.13E+00	6.15E-01	1.27E+00	0.499573	
factor(INHSAMHH)41	2.91E-01	1.34E+00	7.48E-01	9.11E-01	1.96E+00	0.13734	
factor(INHSAMHH)42	-5.57E-01	5.73E-01	1.75E+00	2.12E-01	1.55E+00	0.272837	
factor(INHSAMHH)43	-6.41E-02	9.38E-01	1.07E+00	7.11E-01	1.24E+00	0.650629	
factor(INHSAMHH)51	-2.02E-01	8.17E-01	1.22E+00	4.72E-01	1.41E+00	0.470594	
factor(INHSAMHH)52	-3.75E-01	6.87E-01	1.46E+00	3.33E-01	1.42E+00	0.310015	
factor(INHSAMHH)53	3.48E-01	1.42E+00	7.06E-01	5.41E-01	3.71E+00	0.478751	
factor(INHSAMHH)54	-7.63E-01	4.66E-01	2.15E+00	1.62E-01	1.34E+00	0.155983	
factor(INHSAMHH)55	-1.24E+01	4.23E-06	2.36E+05	2.01E-06	8.90E-06	< 2E-16	***
factor(INHSAMHH)56	3.76E-01	1.46E+00	6.86E-01	1.57E-01	1.36E+01	0.740856	
factor(INHSAMHH)57	2.76E-01	1.32E+00	7.59E-01	3.87E-01	4.49E+00	0.658884	
factor(INHSAMHH)58	9.26E-02	1.10E+00	9.12E-01	6.88E-01	1.75E+00	0.697312	
factor(INHSAMHH)71	-1.17E+01	8.72E-06	1.15E+05	1.98E-06	3.84E-05	< 2E-16	***
factor(INPPOSHHK)2	1.28E-01	1.14E+00	8.80E-01	8.83E-01	1.46E+00	0.320434	
factor(INPPOSHHK)3	2.12E-01	1.24E+00	8.09E-01	9.62E-01	1.59E+00	0.097218	.
factor(INPPOSHHK)4	1.83E-01	1.20E+00	8.33E-01	8.89E-01	1.62E+00	0.232733	
factor(INPPOSHHK)6	2.82E-01	1.33E+00	7.54E-01	7.86E-01	2.24E+00	0.290608	
factor(INPPOSHHK)7	-4.75E-02	9.54E-01	1.05E+00	5.82E-01	1.56E+00	0.850449	
VEHP100WELVAART	-5.61E-04	9.99E-01	1.00E+00	9.97E-01	1.00E+00	0.679048	
RRSYKA	-1.15E-03	9.99E-01	1.00E+00	9.97E-01	1.00E+00	0.301956	
RRDIKA	2.78E-04	1.00E+00	1.00E+00	9.96E-01	1.00E+00	0.889977	
CHOLB	-7.56E-02	9.27E-01	1.08E+00	8.41E-01	1.02E+00	0.127799	
LDLB	-7.61E-03	9.92E-01	1.01E+00	8.99E-01	1.10E+00	0.879791	
HDLB	-3.13E-02	9.69E-01	1.03E+00	8.56E-01	1.10E+00	0.619834	
TRIGB	3.55E-02	1.04E+00	9.65E-01	9.96E-01	1.08E+00	0.079764	.
QUETAO	-8.69E-04	9.99E-01	1.00E+00	9.93E-01	1.01E+00	0.784727	
factor(Armoedegrens)1	-3.67E-01	6.93E-01	1.44E+00	5.79E-01	8.29E-01	6.04E-05	***
factor(Gest Besteedbaar Inkomen)2	2.24E-02	1.02E+00	9.78E-01	8.51E-01	1.23E+00	0.811088	
factor(Gest Besteedbaar Inkomen)3	1.45E-01	1.16E+00	8.65E-01	9.06E-01	1.48E+00	0.242782	
factor(SMOKING STATUS)2	5.41E-02	1.06E+00	9.47E-01	9.75E-01	1.14E+00	0.18425	
factor(MEDICATION LABEL)1	1.78E+00	5.92E+00	1.69E-01	5.27E+00	6.64E+00	< 2E-16	***
factor(GBAGESLACHT)2	-4.04E-01	6.68E-01	1.50E+00	6.00E-01	7.43E-01	1.44E-13	***
factor(Herkomst gehercodeerd)1	-1.91E-03	9.98E-01	1.00E+00	8.60E-01	1.16E+00	0.979989	
factor(Herkomst gehercodeerd)2	1.91E-01	1.21E+00	8.26E-01	9.41E-01	1.56E+00	0.136361	
factor(Herkomst gehercodeerd)3	-5.67E-01	5.67E-01	1.76E+00	4.15E-01	7.74E-01	0.000356	***
factor(Herkomst gehercodeerd)4	-9.75E-02	9.07E-01	1.10E+00	6.87E-01	1.20E+00	0.49241	
factor(Herkomst gehercodeerd)5	6.57E-02	1.07E+00	9.36E-01	9.43E-01	1.21E+00	0.298649	
factor(Herkomst gehercodeerd)6	-4.46E-01	6.40E-01	1.56E+00	3.19E-01	1.29E+00	0.209733	
factor(Herkomst gehercodeerd)7	-1.80E-02	9.82E-01	1.02E+00	8.87E-01	1.09E+00	0.729049	

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.749 (se = 0.004)
Likelihood ratio test= 2427 on 63 df, p=<2e-16
Wald test = 7510 on 63 df, p=<2e-16
Score (logrank) test = 2230 on 63 df, p=<2e-16, Robust = 1725 p=<2e-16
(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

Table B.4: Schoenfeld residuals test of actual Cox model (before adding interaction)

Variable name	rho	chisq	p
MEDICATION COUNT	0.057796	11.77863	0.000599
INHARMLAG	0.014053	0.980658	0.322036
factor(INHBBIHJ)12	-0.01512	0.889199	0.345694
factor(INHBBIHJ)13	-0.01566	0.988668	0.320068
factor(INHBBIHJ)14	-0.00108	0.004262	0.947946
factor(INHBBIHJ)21	-0.0011	0.004705	0.945313
factor(INHBBIHJ)22	-0.0315	4.142312	0.041824
factor(INHBBIHJ)23	0.005043	0.099052	0.75297
factor(INHBBIHJ)24	0.010724	0.454298	0.500301
factor(INHBBIHJ)25	-0.00556	0.116423	0.732947
factor(INHBBIHJ)26	-0.06945	2.163149	0.141355
factor(INHBBIHJ)30	0.037652	4.804707	0.028382
INHP100HBEST	-0.03672	5.869428	0.015406
factor(INHPOPIIV)2	-0.05146	1.021477	0.312169
factor(INHPOPIIV)3	0.025106	0.209803	0.646922
factor(INHPOPIIV)7	-0.03109	3.689553	0.054754
factor(INHPOPIIV)8	0.009629	0.232031	0.630022
factor(INHSAMHH)12	0.005121	0.104792	0.746153
factor(INHSAMHH)13	-0.00401	0.061096	0.804772
factor(INHSAMHH)14	0.014659	0.83585	0.360586
factor(INHSAMHH)21	-0.00025	0.000227	0.987972
factor(INHSAMHH)22	0.001773	0.010734	0.917482
factor(INHSAMHH)31	-0.00875	0.297649	0.58536
factor(INHSAMHH)32	-0.00292	0.030399	0.861588
factor(INHSAMHH)33	-0.00269	0.025311	0.873596
factor(INHSAMHH)41	0.007836	0.228638	0.632536
factor(INHSAMHH)42	0.004109	0.068785	0.793114
factor(INHSAMHH)43	0.015385	0.943255	0.331442
factor(INHSAMHH)51	-0.0104	0.382696	0.536164
factor(INHSAMHH)52	-0.00609	0.137631	0.710647
factor(INHSAMHH)53	-0.00678	0.160513	0.688685
factor(INHSAMHH)54	0.012011	0.562424	0.453285
factor(INHSAMHH)55	-0.00477	0.027457	0.868392
factor(INHSAMHH)56	0.003806	0.06788	0.794449
factor(INHSAMHH)57	-0.01278	0.683987	0.408217
factor(INHSAMHH)58	0.011491	0.398852	0.527683
factor(INHSAMHH)71	-0.03261	0.309159	0.578197
factor(INPPOSHHK)2	0.013979	0.652177	0.419335
factor(INPPOSHHK)3	0.022085	1.608748	0.204668
factor(INPPOSHHK)4	0.010866	0.41223	0.52084
factor(INPPOSHHK)6	0.003207	0.038607	0.84423
factor(INPPOSHHK)7	0.02063	1.485237	0.222957
VEHP100WELVAART	0.048549	9.005178	0.002692
RRSYKA	-0.01342	0.766816	0.381204
RRDIKA	0.010927	0.533061	0.465323
CHOLB	0.002651	0.017268	0.895451
LDLB	-0.0028	0.019141	0.889962
HDLB	0.019906	1.181591	0.277032
TRIGB	-0.01361	0.360092	0.548455
QUETAO	0.027778	2.632055	0.104726
factor(Armoedegrens)1	-0.06036	14.51935	0.000139
factor(Gest Besteedbaar Inkomen)2	-0.00652	0.151628	0.696985
factor(Gest Besteedbaar Inkomen)3	0.00824	0.246007	0.619901
factor(SMOKING STATUS)2	-0.04134	6.3991	0.011418
factor(MEDICATION LABEL)1	-0.19732	232.586	1.63E-52
factor(GBAGESLACHT)2	-0.02076	1.608447	0.20471
factor(Herkomst gehercodeerd)1	0.006619	0.181629	0.669977
factor(Herkomst gehercodeerd)2	0.012484	0.632568	0.426415
factor(Herkomst gehercodeerd)3	0.010612	0.438385	0.507903
factor(Herkomst gehercodeerd)4	0.032631	4.14476	0.041764
factor(Herkomst gehercodeerd)5	0.02041	1.561625	0.211427
factor(Herkomst gehercodeerd)6	0.045926	7.247509	0.0071
factor(Herkomst gehercodeerd)7	0.041388	6.672765	0.00979
GLOBAL	#N/A	334.7279	1.14E-38

Table B.5: Actual Cox Model (after adding interaction)

Variable name	coef	exp(coef)	exp(-coef)	lower .95	upper .95	Pr(>|z|)	
INHARMLAG	-9.248E-04	9.991E-01	1.00E+00	9.99E-01	1.00E+00	3.038E-03	**
factor(INHPOPIIV)2	-1.153E+01	9.788E-06	1.02E+05	4.10E-06	2.34E-05	< 2E-16	***
factor(INHPOPIIV)3	1.099E+01	5.934E+04	1.69E-05	1.29E+04	2.73E+05	< 2E-16	***
factor(INHPOPIIV)7	-4.655E-01	6.278E-01	1.59E+00	2.60E-01	1.52E+00	3.012E-01	
factor(INHPOPIIV)8	1.287E+01	3.870E+05	2.58E-06	3.40E+04	4.41E+06	< 2E-16	***
factor(INHSAMHH)12	-3.360E-01	7.146E-01	1.40E+00	5.58E-01	9.15E-01	7.594E-03	**
factor(INHSAMHH)13	7.077E-02	1.073E+00	9.32E-01	8.54E-01	1.35E+00	5.434E-01	
factor(INHSAMHH)14	-1.266E-01	8.811E-01	1.14E+00	6.80E-01	1.14E+00	3.371E-01	
factor(INHSAMHH)21	-7.043E-02	9.320E-01	1.07E+00	6.57E-01	1.32E+00	6.936E-01	
factor(INHSAMHH)22	-2.924E-01	7.464E-01	1.34E+00	5.15E-01	1.08E+00	1.235E-01	
factor(INHSAMHH)31	-1.081E-01	8.976E-01	1.11E+00	5.94E-01	1.36E+00	6.082E-01	
factor(INHSAMHH)32	-5.122E-01	5.992E-01	1.67E+00	3.55E-01	1.01E+00	5.468E-02	.
factor(INHSAMHH)33	-6.648E-02	9.357E-01	1.07E+00	6.26E-01	1.40E+00	7.461E-01	
factor(INHSAMHH)41	1.820E-01	1.200E+00	8.34E-01	8.09E-01	1.78E+00	3.661E-01	
factor(INHSAMHH)42	-6.980E-01	4.976E-01	2.01E+00	1.82E-01	1.36E+00	1.731E-01	
factor(INHSAMHH)43	-8.401E-02	9.194E-01	1.09E+00	6.92E-01	1.22E+00	5.630E-01	
factor(INHSAMHH)51	-1.688E-01	8.447E-01	1.18E+00	4.76E-01	1.50E+00	5.644E-01	
factor(INHSAMHH)52	-4.468E-01	6.397E-01	1.56E+00	2.95E-01	1.39E+00	2.575E-01	
factor(INHSAMHH)53	3.433E-01	1.410E+00	7.09E-01	5.20E-01	3.82E+00	4.998E-01	
factor(INHSAMHH)54	-6.426E-01	5.259E-01	1.90E+00	1.81E-01	1.53E+00	2.367E-01	
factor(INHSAMHH)55	-1.230E+01	4.565E-06	2.19E+05	2.22E-06	9.38E-06	< 2E-16	***
factor(INHSAMHH)56	2.203E-01	1.246E+00	8.02E-01	1.35E-01	1.15E+01	8.462E-01	
factor(INHSAMHH)57	1.106E-01	1.117E+00	8.95E-01	3.32E-01	3.75E+00	8.580E-01	
factor(INHSAMHH)58	4.203E-02	1.043E+00	9.59E-01	6.45E-01	1.69E+00	8.639E-01	
factor(INHSAMHH)71	-1.156E+01	9.571E-06	1.05E+05	2.10E-06	4.37E-05	< 2E-16	***
factor(INPPOSHHK)2	-1.797E-02	9.822E-01	1.02E+00	7.38E-01	1.31E+00	9.020E-01	
factor(INPPOSHHK)3	7.408E-02	1.077E+00	9.29E-01	8.12E-01	1.43E+00	6.078E-01	
factor(INPPOSHHK)4	5.360E-02	1.055E+00	9.48E-01	7.60E-01	1.47E+00	7.492E-01	
factor(INPPOSHHK)6	1.719E-01	1.188E+00	8.42E-01	6.97E-01	2.02E+00	5.276E-01	
factor(INPPOSHHK)7	2.273E-03	1.002E+00	9.98E-01	6.08E-01	1.65E+00	9.929E-01	
RRSYKA	-1.020E-03	9.990E-01	1.00E+00	9.97E-01	1.00E+00	3.598E-01	
RRDIKA	1.076E-04	1.000E+00	1.00E+00	9.96E-01	1.00E+00	9.572E-01	
CHOLB	-5.669E-02	9.449E-01	1.06E+00	8.58E-01	1.04E+00	2.499E-01	
LDLB	-2.958E-02	9.709E-01	1.03E+00	8.80E-01	1.07E+00	5.541E-01	
HDLB	-5.820E-02	9.435E-01	1.06E+00	8.34E-01	1.07E+00	3.550E-01	
TRIGB	2.621E-02	1.027E+00	9.74E-01	9.85E-01	1.07E+00	2.091E-01	
QUETAO	-1.291E-03	9.987E-01	1.00E+00	9.93E-01	1.01E+00	6.843E-01	
factor(Armoedegrens)1	5.421E-01	1.720E+00	5.82E-01	1.01E+00	2.93E+00	4.555E-02	*
AGE AT BASELINE	-2.370E-02	9.766E-01	1.02E+00	9.57E-01	9.97E-01	2.226E-02	*
factor(INHBBIHJ)12	3.870E-01	1.473E+00	6.79E-01	4.23E-01	5.12E+00	5.428E-01	
factor(INHBBIHJ)13	1.139E-01	1.121E+00	8.92E-01	3.92E-01	3.20E+00	8.315E-01	
factor(INHBBIHJ)14	1.279E+00	3.594E+00	2.78E-01	7.93E-02	1.63E+02	5.109E-01	
factor(INHBBIHJ)21	-8.854E-01	4.125E-01	2.42E+00	8.90E-02	1.91E+00	2.580E-01	
factor(INHBBIHJ)22	7.296E-01	2.074E+00	4.82E-01	7.17E-01	6.00E+00	1.780E-01	
factor(INHBBIHJ)23	-1.105E+00	3.311E-01	3.02E+00	8.08E-02	1.36E+00	1.246E-01	
factor(INHBBIHJ)24	1.899E-01	1.209E+00	8.27E-01	4.63E-01	3.16E+00	6.981E-01	
factor(INHBBIHJ)25	-4.329E-01	6.486E-01	1.54E+00	2.98E-01	1.41E+00	2.763E-01	
factor(INHBBIHJ)26	-1.417E+01	7.008E-07	1.43E+06	1.42E-08	3.47E-05	1.090E-12	***
factor(INHBBIHJ)30	-2.145E+00	1.170E-01	8.55E+00	2.25E-02	6.08E-01	1.068E-02	*
factor(Gest Besteedbaar Inkomen)2	5.950E-02	1.061E+00	9.42E-01	8.82E-01	1.28E+00	5.289E-01	
factor(Gest Besteedbaar Inkomen)3	2.212E-01	1.248E+00	8.02E-01	9.75E-01	1.60E+00	7.861E-02	.
INHP100HBEST	1.385E-02	1.014E+00	9.86E-01	1.00E+00	1.03E+00	5.862E-02	.
VEHP100WELVAART	-2.059E-02	9.796E-01	1.02E+00	9.68E-01	9.91E-01	7.170E-04	***
factor(SMOKING STATUS)2	4.363E-01	1.547E+00	6.46E-01	1.12E+00	2.13E+00	7.835E-03	**
factor(MEDICATION LABEL)1	3.913E+00	5.004E+01	2.00E-02	3.36E+01	7.46E+01	< 2E-16	***
MEDICATION COUNT	-1.251E-02	9.876E-01	1.01E+00	9.53E-01	1.02E+00	4.968E-01	
factor(GBAGESLACHT)2	-4.202E-01	6.569E-01	1.52E+00	5.89E-01	7.33E-01	3.830E-14	***
factor(Armoedegrens)1 :AGE AT BASELINE	-1.632E-02	9.838E-01	1.02E+00	9.75E-01	9.93E-01	3.010E-04	***
factor(INHBBIHJ)12 :AGE AT BASELINE	-7.182E-03	9.928E-01	1.01E+00	9.67E-01	1.02E+00	5.929E-01	
factor(INHBBIHJ)13 :AGE AT BASELINE	-2.414E-03	9.976E-01	1.00E+00	9.75E-01	1.02E+00	8.347E-01	
factor(INHBBIHJ)14 :AGE AT BASELINE	-2.683E-02	9.735E-01	1.03E+00	8.94E-01	1.06E+00	5.364E-01	
factor(INHBBIHJ)21 :AGE AT BASELINE	1.526E-02	1.015E+00	9.85E-01	9.83E-01	1.05E+00	3.503E-01	
factor(INHBBIHJ)22 :AGE AT BASELINE	-1.644E-02	9.837E-01	1.02E+00	9.61E-01	1.01E+00	1.620E-01	
factor(INHBBIHJ)23 :AGE AT BASELINE	1.520E-02	1.015E+00	9.85E-01	9.90E-01	1.04E+00	2.409E-01	
factor(INHBBIHJ)24 :AGE AT BASELINE	3.765E-03	1.004E+00	9.96E-01	9.83E-01	1.03E+00	7.186E-01	
factor(INHBBIHJ)25 :AGE AT BASELINE	7.495E-03	1.008E+00	9.93E-01	9.94E-01	1.02E+00	2.939E-01	
factor(INHBBIHJ)26 :AGE AT BASELINE	5.560E-02	1.057E+00	9.46E-01	9.60E-01	1.16E+00	2.576E-01	
factor(INHBBIHJ)30 :AGE AT BASELINE	3.096E-02	1.031E+00	9.70E-01	1.01E+00	1.06E+00	1.336E-02	*
INHP100HBEST :AGE AT BASELINE	-1.840E-04	9.998E-01	1.00E+00	1.00E+00	1.00E+00	9.500E-02	.
VEHP100WELVAART :AGE AT BASELINE	3.138E-04	1.000E+00	1.00E+00	1.00E+00	1.00E+00	6.930E-04	***
factor(SMOKING STATUS)2 :AGE AT BASELINE	-6.957E-03	9.931E-01	1.01E+00	9.88E-01	9.99E-01	1.521E-02	*
factor(MEDICATION LABEL)1 :AGE AT BASELINE	-3.794E-02	9.628E-01	1.04E+00	9.57E-01	9.69E-01	< 2E-16	***
MEDICATION COUNT :AGE AT BASELINE	4.457E-04	1.000E+00	1.00E+00	1.00E+00	1.00E+00	1.469E-01	

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.745 (se = 0.005)
Likelihood ratio test= 2632 on 73 df, p=<2e-16
Wald test = 6330 on 73 df, p=<2e-16
Score (logrank) test = 2668 on 73 df, p=<2e-16, Robust = 1745 p=<2e-16
(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

Table B.6: Schoenfeld residuals test of actual Cox model (after adding interaction)

Variable name	rho	chisq	p
INHARMLAG	0.014431	1.07247	0.300388
factor(INHPOPIIV)2	-0.00308	0.004851	0.944475
factor(INHPOPIIV)3	0.004644	0.009792	0.921175
factor(INHPOPIIV)7	-0.00767	0.21745	0.64099
factor(INHPOPIIV)8	0.008192	0.163628	0.685838
factor(INHSAMHH)12	0.006487	0.169855	0.680241
factor(INHSAMHH)13	0.003283	0.04197	0.837678
factor(INHSAMHH)14	0.011913	0.567386	0.4513
factor(INHSAMHH)21	-0.01052	0.463461	0.496011
factor(INHSAMHH)22	-0.00721	0.205615	0.650226
factor(INHSAMHH)31	-0.0109	0.519106	0.471223
factor(INHSAMHH)32	-0.00229	0.020704	0.885589
factor(INHSAMHH)33	-0.00931	0.353036	0.5524
factor(INHSAMHH)41	0.003385	0.045336	0.831387
factor(INHSAMHH)42	0.001686	0.011904	0.91312
factor(INHSAMHH)43	0.014233	0.807605	0.368829
factor(INHSAMHH)51	-0.01528	0.867277	0.35171
factor(INHSAMHH)52	-0.00353	0.052806	0.818251
factor(INHSAMHH)53	0.001201	0.005162	0.942726
factor(INHSAMHH)54	0.006884	0.185959	0.666302
factor(INHSAMHH)55	-0.01173	0.199085	0.65546
factor(INHSAMHH)56	-0.00343	0.057309	0.810801
factor(INHSAMHH)57	-0.01129	0.517915	0.471732
factor(INHSAMHH)58	0.010134	0.326703	0.567607
factor(INHSAMHH)71	-0.0107	0.047689	0.827134
factor(INPPOSHHK)2	0.022344	2.103354	0.146976
factor(INPPOSHHK)3	0.026201	2.856534	0.091003
factor(INPPOSHHK)4	0.01793	1.329295	0.248931
factor(INPPOSHHK)6	0.00481	0.092022	0.761622
factor(INPPOSHHK)7	0.015293	0.875957	0.349311
RRSYKA	-0.01051	0.470356	0.492823
RRDIKA	0.006233	0.173549	0.676977
CHOLB	0.003269	0.025359	0.873476
LDLB	-0.00311	0.023187	0.878973
HDLB	0.019552	1.126073	0.288615
TRIGB	-0.0087	0.159262	0.689837
QUETAO	0.024565	2.034563	0.153758
factor(Armoedegrens)1	0.00161	0.012015	0.912715
AGE AT BASELINE ADJ	-0.01519	0.973021	0.323928
factor(INHBBIHJ)12	0.035277	1.995841	0.157731
factor(INHBBIHJ)13	-0.01412	1.009424	0.315041
factor(INHBBIHJ)14	-0.01135	1.676643	0.195371
factor(INHBBIHJ)21	0.014796	0.518562	0.471456
factor(INHBBIHJ)22	-0.01381	1.030259	0.310098
factor(INHBBIHJ)23	-0.00224	0.01962	0.888605
factor(INHBBIHJ)24	0.006827	0.165123	0.684483
factor(INHBBIHJ)25	0.006643	0.202995	0.652314
factor(INHBBIHJ)26	0.005044	0.011221	0.915637
factor(INHBBIHJ)30	0.01355	0.635727	0.425262
factor(Gest Besteedbaar Inkomen)2	0.005975	0.128417	0.720079
factor(Gest Besteedbaar Inkomen)3	0.012941	0.616877	0.43221
INHP100HBEST	-0.02622	3.067236	0.079885
VEHP100WELVAART	0.029216	3.683812	0.054943
factor(SMOKING STATUS)2	0.029635	3.44406	0.06348
factor(MEDICATION LABEL)1	0.001688	0.013614	0.907115
MEDICATION COUNT	-0.02035	1.464613	0.226198
factor(GBAGESLACHT)2	-0.01726	1.138142	0.286045
factor(Armoedegrens)1 :AGE AT BASELINE	-0.00656	0.196844	0.657281
factor(INHBBIHJ)12 :AGE AT BASELINE	-0.04169	2.691973	0.100855
factor(INHBBIHJ)13 :AGE AT BASELINE	0.013307	0.898911	0.343074
factor(INHBBIHJ)14 :AGE AT BASELINE	0.011629	1.772008	0.183134
factor(INHBBIHJ)21 :AGE AT BASELINE	-0.01758	0.685811	0.407593
factor(INHBBIHJ)22 :AGE AT BASELINE	0.014293	1.08097	0.298481
factor(INHBBIHJ)23 :AGE AT BASELINE	0.002207	0.017866	0.893668
factor(INHBBIHJ)24 :AGE AT BASELINE	-0.00787	0.215761	0.642289
factor(INHBBIHJ)25 :AGE AT BASELINE	-0.00749	0.261228	0.609278
factor(INHBBIHJ)26 :AGE AT BASELINE	-0.00864	0.026514	0.87065
factor(INHBBIHJ)30 :AGE AT BASELINE	-0.01275	0.588849	0.442865
INHP100HBEST :AGE AT BASELINE	0.0225	2.104957	0.146822
VEHP100WELVAART :AGE AT BASELINE	-0.03017	3.762215	0.052423
factor(SMOKING STATUS)2 :AGE AT BASELINE	-0.03176	3.784617	0.051726
factor(MEDICATION LABEL)1 :AGE AT BASELINE	-0.00594	0.163072	0.686344
MEDICATION COUNT :AGE AT BASELINE	0.021683	1.588052	0.207605
GLOBAL	#N/A	43.47484	0.997637

Table B.7: Cox model interpretation part 1

Variable	Exp(coef) (SD)	Interpretation
INHARMLAG	0.9991 (0.9985 - 0.9997)	Holding the other value constant, having higher the INHARMLAG (income according to poverty line) relate to good prognostic by factor 0.9991 or 0.009%
(INHPOPIIV)2	0.000009788 (0.0000041 - 0.00002337)	Holding the other value constant, INHPOPIIV is 2 (student household with income), relate to good prognostic by factor 0.000009788 or 99.9%. However this number is really small, and there is a possibility of overfitting because of the censored to event ratio
(INHPOPIIV)3	59340 (12880 - 273400)	Holding the other value constant, INHPOPIIV is 3 (institution household with income), relate to bad prognostic by factor 59340. However this number is really big, and there is a possibility of overfitting because of the censored to event ratio
(INHPOPIIV)8	387000 (33960 - 4411000)	Holding the other value constant, INHPOPIIV 8 (institution household without income), relate to bad prognostic by factor 387000. However this number is really big, and there is a possibility of overfitting because of the censored to event ratio
(INHSAMHH)12	0.7146 (0.5584 - 0.9146)	Holding the other value constant, INHSAMHH 12 (Single man from AOW), relate to good prognostic by factor 0. 7146 or by 28.54%.
(INHSAMHH)55	0.000004565 (0.000002221 - 0.000009383)	Holding the other value constant, INHSAMHH is 55 (Single parent with minor child and other resident), relate to good prognostic by factor 0.000004565 or 99.9%. However this number is really small, and there is a possibility of overfitting because of the censored to event ratio
(INHSAMHH)71	0.000009571 (0.000002099 - 0.00004365)	Holding the other value constant, INHSAMHH 71 (population in institution, institution and homes), relate to good prognostic by factor 0.000009571 or 99.9%. However this number is really small, and there is a possibility of overfitting because of the censored to event ratio
(INHBBIHJ)26	0.0000007008 (0.00000001416 - 0.00003469)	Holding the other value constant, INHBBIHJ 26 (Student finance), relate to good prognostic by factor 0.0000007008 or 99.9%. However this number is really small, and there is a possibility of overfitting because of the censored to event ratio
(GBAGESLACHT)2	0.6569 (0.5892 - 0.7325)	Holding the other value constant, having GBAGESLACHT is 2, relate to good prognostic by factor 0.6569

Table B.8: Cox model interpretation part 2

Variable	Exp(coef) (SD)	Interpretation
AGE AT BASELINE	0.9766 (0.9569 - 0.9966)	The effect of age need to be adjusted for all the interaction, therefore the summation of all variables that have interaction with this variable need to be calculated.
(Armoedegrens)1	1.72 (1.011 - 2.925)	Holding the other value constant, Armoedegrens is 1 (Above poverty line), relate to bad prognostic by factor 1. 72.
(Armoedegrens)1:AGE AT BASELINE	0.9838 (0.9751 - 0.9926)	Holding the other value constant, given the higher AGE AT BASELINE , having Armoedegrens 1 (above poverty line), relate to good prognostic by factor 0.9838.
(INHBBIHJ)30	0.117 (0.02254 - 0.6075)	Holding the other value constant, having INHBBIHJ is 30 (Property income), relate to good prognostic by factor 0.117.
(INHBBIHJ)30:AGE AT BASELINE	1.031 (1.006 - 1.057)	Holding the other value constant, given the higher AGE AT BASELINE , having INHBBIHJ 30 (Property income), relate to bad prognostic by factor 1.031.
VEHP100WELVAART	0.9796 (0.968 - 0.9914)	Holding the other value constant, having higher the VEHP100WELVAART (Prosperity : Income + wealth), relate to good prognostic by factor 0.9796 or 2.4%
VEHP100WELVAART:AGE AT BASELINE	1.0003 (1 – 1)	Holding the other value constant, given the higher AGE AT BASELINE , having higher VEHP100WELVAART (Prosperity : Income + wealth), relate to bad prognostic by factor 1.0003.
(SMOKING STATUS)2	1.547 (1.122 - 2.134)	Holding the other value constant, having SMOKING STATUS 2 (smoking people), relate to bad prognostic by factor 1.547
(SMOKING STATUS)2:AGE AT BASELINE	0.9931 (0.9875 - 0.9987)	Holding the other value constant, given the higher AGE AT BASELINE , having SMOKING STATUS 2 (smoking people), relate to good prognostic by factor 0.9931.
(MEDICATION LABEL)1	50.04 (33.58 – 74.56)	Holding the other value constant, having MEDICATION LABEL 1, relate to bad prognostic by factor 50.04. However this number is really big, and there is a possibility of overfitting because of the censored to event ratio
(MEDICATION LABEL)1:AGE AT BASELINE	0.9628 (0.9565 - 0.9690)	Holding the other value constant, given the higher AGE AT BASELINE , having MEDICATION LABEL 1 (use CVD treatment), relate to good prognostic by factor 0.9628.

B.3. Bayesian modelling result

Table B.9: Discrete Bayesian Network Result for AGE

AGE_G	MLE hazards rate	Bayesian hazards rate
25-30	0.000	0.120
30-35	0.001	0.014
35-40	0.001	0.010
40-45	0.002	0.012
45-50	0.004	0.014
50-55	0.006	0.018
55-60	0.008	0.023
60-65	0.010	0.027
65-70	0.012	0.031
70-75	0.016	0.043
75-80	0.019	0.056
80-85	0.028	0.073
85-90	0.037	0.117
90-95	0.068	0.173
95-100	0.063	0.255
100+	0.100	0.440

Table B.10: Discrete Bayesian Network result for GBAGESLACHT

SEX_G	MLE hazards rate	Bayesian hazards rate
FEMALE	0.008278602	0.027573262
MALE	0.011085142	0.034334264

Table B.11: Discrete Bayesian Network result for Herkomst gehercodeerd

SUB_ETHNICITY_G	MLE hazards rate	Bayesian hazards rate
Antillen and Aruba	0.010259658	0.120717228
Dutch	0.009755448	0.01330426
Indonesian	0.008561854	0.043968978
Marokoans	0.005653977	0.121818535
Other	0.009894234	0.025630733
Polish	0.003529412	0.261017838
Surinamese	0.007744176	0.056756607
Turkish	0.01173898	0.140139688

Table B.12: Discrete Bayesian Network result for Armoedegrens

POVERTY_G	MLE hazards rate	Bayesian hazards rate
Above Poverty	0.008851	0.022486
Under Poverty	0.015381	0.093734

Table B.13: Discrete Bayesian Network result for Gest Besteedbaar Inkomen

INCOME_G	MLE hazards rate	Bayesian hazards rate
High	0.007766	0.023416
Middle	0.009816	0.029404
Low	0.011279	0.069354

Table B.14: Discrete Bayesian Network result for Welvaart

PROSPERITY_G	MLE hazards rate	Bayesian hazards rate
High	0.01162	0.0287
Middle	0.00920	0.0278
Low	0.00931	0.0668

Table B.15: Discrete Bayesian Network result for INHBBIHJ

SOURCE_INCOME_G	MLE hazards rate	Bayesian hazards rate
Other income self-employed	0.005327	0.065931
Other social security benefits	0.008198	0.051668
Profit self-employed	0.005184	0.022313
Property income	0.017124	0.060604
Retirement Benefit	0.018467	0.051176
Salary	0.00466	0.014036
Shareholder	0.005971	0.017658
Sickness benefit	0.007483	0.036586
Social assistance benefit	0.007572	0.057254
Study grants	0.010989	0.276532
Unemployment benefits	0.006453	0.035928

Table B.16: Discrete Bayesian Network result for INHSAMHH

MARTIAL_STATUS_G	MLE hazards rate	Bayesian hazards rate
Couple with adult kids	0.00776	0.019976
Couple with adult kids and other resident	0.006606	0.018969
Couple with minor and adult kids	0.004964	0.013402
Couple with minor and adult kids and other resident	0.005764	0.022222
Couple with minor kids	0.002669	0.013359
Couple with minor kids and other resident	0.005705	0.018087
Couple without children	0.006302	0.018443
Couple without children AOW	0.016962	0.045319
Couple without children with other resident	0.006736	0.024009
Other personal household	0.005799	0.022151
Population institution household	0.050524	0.155613
Single man	0.006504	0.024689
Single man AOW	0.0236	0.074205
Single with adult kids	0.00787	0.025381
Single with adult kids and other resident	0.003503	0.024896
Single with minor and adult kids	0.004923	0.017692
Single with minor and adult kids and other resident	0.017857	0.020548
Single with minor kids	0.002671	0.019093
Single with minor kids and other resident	0.001946	0.013804
Single woman	0.005243	0.02313
Single woman AOW	0.018282	0.054936

Table B.17: Discrete Bayesian Network result for SMOKING STATUS

SMOKING_STATUS_G	MLE hazards rate	Bayesian hazards rate
Yes	0.007367	0.037381
No	0.0105	0.028218

Table B.18: Discrete Bayesian Network result for MEDICATION LABEL

MEDICATION_LABEL_G	MLE hazards rate	Bayesian hazards rate
Not treated	0.002413	0.019475
Treated	0.016976	0.04123

Bibliography

Abeysekera, W. and Sooriyarachchi, R. (2009). Use of schoenfeld's global test to test the proportional hazards assumption in the cox proportional hazards model: an application to a clinical study. *Journal of the National Science Foundation of Sri Lanka*.

Akhil, M., Chandra, P., and Deekshatulu, B. L. (2012). Heart Disease Prediction System using Associative Classification and Genetic Algorithm. *Icecit 2012*.

Assmann, G., Cullen, P., and Schulte, H. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular münster (PROCAM) study. *Circulation*, 105(3):310–315.

Ata, N. and Sözer, M. T. (2007). Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36(2):157–167.

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E., and O'Connor, P. J. (2015). Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29(4):1033–1069.

Beltrán-Sánchez, H., Crimmins, E. M., and Finch, C. E. (2012). Early cohort mortality predicts the rate of aging in the cohort: A historical analysis. *Journal of Developmental Origins of Health and Disease*, 3(5):380–386.

Birtcher, K. K. and Ballantyne, C. M. (2004). Measurement of cholesterol: a patient perspective. *Circulation*, 110(11):e296–e297.

Bittl, J. A. and He, Y. (2017). Bayesian Analysis: A Practical Approach to Interpret Clinical Trials and Create Clinical Practice Guidelines. *Circulation: Cardiovascular Quality and Outcomes*, 10(8):1–11.

Bos, V., Kunst, A. E., Keij-Deerenberg, I. M., Garssen, J., and Mackenbach, J. P. (2004). Ethnic inequalities in age- and cause-specific mortality in The Netherlands. *International Journal of Epidemiology*, 33(5):1112–1119.

Bradburn, M., Clark, T., Love, S., and Altman, D. (2003a). Survival analysis part iii: multivariate data analysis–choosing a model and assessing its adequacy and fit. *British journal of cancer*, 89(4):605.

Bradburn, M. J., Clark, T. G., Love, S., and Altman, D. (2003b). Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431.

Cain, K. C., Harlow, S. D., Little, R. J., Nan, B., Yosef, M., Taffe, J. R., and Elliott, M. R. (2011). Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *American journal of epidemiology*, 173(9):1078–1084.

Canchola, A. J., Stewart, S. L., Bernstein, L., West, D. W., Ross, R. K., Deapen, D., Pinder, R., Reynolds, P., Wright, W., Anton-Culver, H., et al. (2003). Cox regression using different time-scales. *Western Users of SAS Software. San Francisco, California*.

CBS (2018a). `https://opendata.cbs.nl/statline/#/CBS/en/dataset/7052eng/table?ts=1542731085077`.

CBS (2018b). Catalogus microdata. `https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/catalogus-microdata`.

Checkland, K. (2004). National service frameworks and uk general practitioners: street-level bureaucrats at work? *Sociology of health & illness*, 26(7):951–975.

Cheung, Y. B., Gao, F., and Khoo, K. S. (2003). Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *Journal of clinical epidemiology*, 56(1):38–43.

Clark, T., Bradburn, M., Love, S., and Altman, D. (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232.

Conroy, R. M., Pyörälä, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetière, P., Jousilahti, P., Keil, U., Njølstad, I., Oganov, R. G., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Witheimsen, L., and Graham, I. M. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *European Heart Journal*, 24(11):987–1003.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*, 117(6):743–753.

Dekker, F. W., De Mutsert, R., Van Dijk, P. C., Zoccali, C., and Jager, K. J. (2008). Survival analysis: time-dependent effects and time-varying risk factors. *Kidney international*, 74(8):994–997.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087 – 1091.

Dunkler, D., Plischke, M., Leffondré, K., and Heinze, G. (2014). Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PLoS One*, 9(11):e113677.

Erasmus, E. (2015). Street-level bureaucracy. *Retrieved September*, 15:2017.

Exter, A. d., Hermans, H., Dosljak, M., Busse, R., Ginneken, E. v., Schreyoegg, J., Wisbaum, W., Organization, W. H., et al. (2004). Health care systems in transition: Netherlands. Technical report, Copenhagen: WHO Regional Office for Europe.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.

for Disease Control, C., Prevention, et al. (2010). How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general. *Centers for Disease Control and Prevention*.

Galobardes, B., Shaw, M., Lawlor, D. A., and Lynch, J. W. (2006a). Indicators of socioeconomic position (part 2). *Journal of epidemiology and community health*, 60(2):95.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., and Smith, G. D. (2006b). Indicators of socioeconomic position (part 1). *Journal of Epidemiology & Community Health*, 60(1):7–12.

Giampaoli, S. (2007). CUORE: A sustainable cardiovascular disease prevention strategy. *European Journal of Preventive Cardiology*, 14(2):161–162.

Goel, M. K., Khanna, P., and Kishore, J. (2010). Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274.

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T., Smith, S. C., Sorlie, P., Stone, N. J., and Wilson, P. W. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 PART B):2935–2959.

Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. A. (2017a). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208.

Goldstein, B. A., Pencina, M. J., Montez-Rath, M. E., and Winkelmayer, W. C. (2017b). Predicting mortality over different time horizons: Which data elements are needed? *Journal of the American Medical Informatics Association*, 24(1):176–181.

Gurrin, L. C., Kurinczuk, J. J., and Burton, P. R. (2000). Bayesian statistics in medical research: An intuitive alternative to conventional data analysis. *Journal of Evaluation in Clinical Practice*, 6(2):193–204.

Hall, P. (2016). A Bayesian Approach to Map-Aided. (AMS).

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hartog, J. and Zorlu, A. (2009). Ethnic segregation in The Netherlands. *International Journal of Manpower*, 30(1/2):15–25.

Hippisley-Cox, J., Coupland, C., Robson, J., and Brindle, P. (2011). Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj*, 342(7788):93.

Hippisley-Cox, J., Coupland, C., Robson, J., Sheikh, A., and Brindle, P. (2009). Predicting risk of type 2 diabetes in England and Wales: Prospective derivation and validation of QDScore. *BMJ (Online)*, 338(7698):811–816.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., and Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *British Medical Journal*, 335(7611):136–141.

Honjo, K. (2004). Social epidemiology: Definition, history, and research examples. *Environmental health and preventive medicine*, 9(5):193–9.

IHCSP (2019). International health care system profiles. https://international.commonwealthfund.org/countries/netherlands/.

Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112.

Kleinert, H. D., Harshfield, G. A., Pickering, T. G., Devereux, R. B., Sullivan, P. A., Marion, R. M., Mallory, W. K., and Laragh, J. H. (1984). What is the value of home blood pressure measurement in patients with mild hypertension? *Hypertension*, 6(4):574–578.

Lamarca, R., Alonso, J., Gomez, G., and Muñoz, Á. (1998). Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 53(5):M337–M343.

Lee, S. J., Boscardin, W. J., Kirby, K. A., and Covinsky, K. E. (2014). Individualizing life expectancy estimates for older adults using the gompertz law of human mortality. *PLoS ONE*, 9(9).

Lewsey, J. D., Lawson, K. D., Ford, I., Fox, K. A. A., Ritchie, L. D., Tunstall-Pedoe, H., Watt, G. C. M., Woodward, M., Kent, S., Neilson, M., and Briggs, A. H. (2015). A cardiovascular disease policy model that predicts life expectancy taking into account socioeconomic deprivation. *Heart*, 101(3):201–208.

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

Lugtenberg, M., Zegers-van Schaick, J. M., Westert, G. P., and Burgers, J. S. (2009). Why don't physicians adhere to guideline recommendations in practice? an analysis of barriers among dutch general practitioners. *Implementation Science*, 4(1):54.

LUMC (2018). Elan-research informatie voor onderzoekers. https://www.lumc.nl/org/pheg/research/collaboration/elan-extramuraal-leids-academisch-netwerk/elan-research-informatie-voor-onderzoekers/.

Markt, A. C. . (2019). Legislation. https://www.acm.nl/en/about-acm/mission-vision-strategy/legislation.

Méjean, C., Droomers, M., van der Schouw, Y. T., Sluijs, I., Czernichow, S., Grobbee, D. E., Bueno-de Mesquita, H. B., and Beulens, J. W. (2013). The contribution of diet and lifestyle to socioeconomic inequalities in cardiovascular morbidity and mortality. *International journal of cardiology*, 168(6):5190–5195.

Miner, L., Bolding, P., Hilbe, J., Goldstein, M., Hill, T., Nisbet, R., Walton, N., and Miner, G. (2014). *Practical predictive analytics and decisioning systems for medicine: Informatics accuracy and cost-effectiveness for healthcare administration and delivery including medical research*. Academic Press.

Nguefack-Tsague, G. (2011). Using Bayesian Networks to Model Hierarchical Relationships in Epidemiological Studies. *Epidemiology and Health*, 33:e2011006.

NHG (2019a). Cardiovasculair risicomanagement. https://www.nhg.org/standaarden/samenvatting/cardiovasculair-risicomanagement/.

NHG (2019b). Tabellen. https://www.nhg.org/themas/artikelen/nhg-tabellen.

Nianogo, R. A. and Arah, O. A. (2015). Agent-based modeling of noncommunicable diseases: A systematic review. *American Journal of Public Health*, 105(3):e20—-e31.

OECD (2018). State of health in the eu : Netherlands - country health profile 2017. http://www.euro.who.int/__data/assets/pdf_file/0005/355991/Health-Profile-Netherlands-Eng.pdf?ua=1.

Overheid (2019a). https://www.ggdhaaglanden.nl/over.htm.

Overheid (2019b). Algemene wet bijzondere ziektekosten. https://wetten.overheid.nl/BWBR0002614/2014-07-16.

Overheid (2019c). Gezondheidswet. https://wetten.overheid.nl/BWBR0002202/2019-02-01.

Overheid (2019d). Wet maatschappelijke ondersteuning 2015. https://wetten.overheid.nl/BWBR0035362/2019-04-02.

Overheid (2019e). Wet publieke gezondheid. https://wetten.overheid.nl/BWBR0024705/2019-01-01.

Overheid (2019f). Wet voorzieningen gehandicapten. https://wetten.overheid.nl/BWBR0006169/2006-03-08.

Pampalon, R., Raymond, G., et al. (2000). A deprivation index for health and welfare planning in quebec. *Chronic Dis Can*, 21(3):104–113.

Park, S. and Hendry, D. J. (2015). Reassessing schoenfeld residual tests of proportional hazards in political science event history analyses. *American Journal of Political Science*, 59(4):1072–1087.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 9:157.

Perini, W., Snijder, M. B., Peters, R. J. G., and Kunst, A. E. (2018). Ethnic disparities in estimated cardiovascular disease risk in Amsterdam, the Netherlands : The HELIUS study. *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*, 26(5):252–262.

Piepoli, M. F., Hoes, A. W., and Agewall, e. a. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, 37(29):2315–2381.

Pignone, M., Phillips, C., and Mulrow, C. (2000). Use of lipid lowering drugs for primary prevention of coronary heart disease: meta-analysis of randomised trials. *Bmj*, 321(7267):983.

Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X., and Eckel, R. H. (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 american heart association scientific statement on obesity and heart disease from the obesity committee of the council on nutrition, physical activity, and metabolism. *Circulation*, 113(6):898–918.

Potvin, L., Richard, L., and Edwards, A. C. (2000). Knowledge of cardiovascular disease risk factors among the canadian population: relationships with indicators of socioeconomic status. *Cmaj*, 162(9 suppl):S5–S11.

Pouwels, K. B., Voorham, J., Hak, E., and Denig, P. (2016). Identification of major cardiovascular events in patients with diabetes using primary care data. *BMC health services research*, 16(1):110.

Psaltopoulou, T., Hatzis, G., Papageorgiou, N., Androulakis, E., Briasoulis, A., and Tousoulis, D. (2017). Socioeconomic status and risk factors for cardiovascular disease: impact of dietary mediators. *Hellenic Journal of Cardiology*, 58(1):32–42.

Rhemtulla, M., Brosseau-Liard, P. É., and Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological methods*, 17(3):354.

RIVM (2019). About rivm. https://www.rivm.nl/en/about-rivm/rivm.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Schäfer, W., Kroneman, M., Boerma, W., Westert, G., Devillé, W., et al. (2010). The netherlands: health system review. *Health systems in transition*, 12(1):v–xxvii.

SCP (2019). Wat is het scp? https://www.scp.nl/Organisatie/Wat_is_het_SCP.

Scutari, M. and Denis, J.-B. (2014). *Bayesian networks: with examples in R*. Chapman and

Hall/CRC.
Server, B. (2017). An introduction to dynamic bayesian networks (dbn). learn how they can be used to model time series and sequences by extending bayesian networks with temporal nodes, allowing prediction into the future, current or past.
Sheldon, T. A. and Parker, H. (1992). Race and ethnicity in health research. *Journal of Public Health*, 14(2):104–110.
Siddique, J. and Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, 27(1):83–102.
Squires, B. P. (2000). Cardiovascular disease and socioeconomic status. *Cmaj*, 162(9 suppl):S3–S3.
Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338.
Therneau, T., Crowson, C., and Atkinson, E. (2013). Using time dependent covariates and time dependent coefficients in the cox model. *Red*, 2:1.
Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., and Ng, M. (2006). Regression for ordinal variables without underlying continuous variables. *Information Sciences*, 176(4):465–474.
Townsend, P. (1987). Deprivation. *Journal of social policy*, 16(2):125–146.
Twardy, C., Nicholson, A., Korb, K., and McNeil, J. (2004). Data mining cardiovascular bayesian networks. *Monash University, School of Computer Science & Software Engineering, Melbourne*, 165.
Twardy, C. R., Nicholson, A. E., Korb, K., and McNeil, J. (2005). Knowledge engineering cardiovascular bayesian networks from the literature. *School of Computer Science & Software Engineering*.
Twardy, C. R., Nicholson, A. E., Korb, K. B., and McNeil, J. (2006). Epidemiological data mining of cardiovascular bayesian networks. *electronic Journal of Health Informatics*, 1(1):3.
Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of clinical Oncology*, 32(22):2380.
Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. (2009). Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363.
Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.
van Gerven, M. A., Taal, B. G., and Lucas, P. J. (2008). Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics*, 41(4):515–529.
van Houwelingen, H. C. and Eilers, P. H. (2000). Non-proportional hazards models in survival analysis. In *COMPSTAT*, pages 151–160. Springer.
Verduijn, M., Peek, N., Rosseel, P. M. J., de Jonge, E., and de Mol, B. A. J. M. (2007). Prognostic Bayesian networks. I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40(6):609–618.
Volksgezondheid, M. v. (2018). Taak. `https://www.gezondheidsraad.nl/over-ons/taak`.
Volksgezondheid, M. v. (2019a). English. `https://www.nza.nl/english`, journal=Nederlandse Zorgautoriteit.
Volksgezondheid, M. v. (2019b). Over de rvs. `https://www.raadrvs.nl/over-de-rvs`, journal=Raad voor Volksgezondheid en Samenleving.
Volksgezondheid, M. v. (2019c). Over ons. `https://www.igj.nl/over-ons`.
Volksgezondheid, M. v. (2019d). Over ons. `https://www.igj.nl/over-ons`, journal=Naar de hompage van igj.nl.
Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):110.
Weatherley, R. and Lipsky, M. (1977). Street-level bureaucrats and institutional innovation: Implementing special-education reform. *Harvard educational review*, 47(2):171–197.
White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equa-

tions: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American sociological review*, pages 512–525.

Woodward, M., Brindle, P., and Tunsfall-Pedoe, H. (2007). Adding social deprivation and family history to cardiovascular risk assessment: The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*, 93(2):172–176.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7).

Zha.nl, Z. (2019). General practitioner. `https://www.ziekenhuisamstelland.nl/en/patients/dutch-health-care-system/general-practitioner/`.