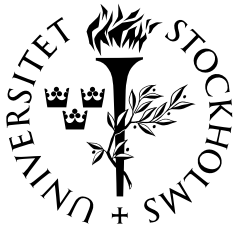


EXAMENSARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Mathematical properties of epidemiological case-cohort designs

av

Karin Fremling

2008 - No 11

Mathematical properties of epidemiological case-cohort designs

Karin Fremling

Examensarbete i matematik 30 högskolepoäng, fördjupningskurs

Handledare: Juni Palmgren
Bihandledare: Samuli Ripatti

2008

Abstract

In this thesis, I describe central concepts in event history analysis, including Cox proportional hazards model, the log-linear model and the illness-death model, and relate them to each other. We are interested in the difference in bias and precision when including, or excluding, the baseline prevalent cases in an analysis of effects of genotype on the hazard using a case-cohort design. I generate populations, according to two models, where the cases, myocardial infarction, depend on the genotype. In one of the models death after MI and prior to baseline also depends on genotype. In the traditional case-cohort analysis only incident new cases during follow-up are included. We enrich the analysis with prevalent cases that are alive at baseline and we expect a selection bias in the association between genotype when death after MI depends on genotype. The results do not, however, indicate any strong selection bias from including prevalent cases in the case-cohort analysis.

Acknowledgment

First I want to thank my supervisor, Juni Palmgren. Some says she is never in her office, that is partly true. But she made some time for me, not always in her office at MEB, but at her office at Kräftriket and in her home. And she is very good at answering emails fast, which I appreciated a lot.

Then I want to thank my co-supervisor, Samuli Ripatti, in Finland with all help with the models and guiding with the software R.

I also want to thank my family and friends, most my father Lennart Fremling.

The biggest thanks goes to my boyfriend, Micke Kardell. He helped me a lot during the time I was working on my thesis. Micke is the one I have discussed most things with, mathematical as well as programming. He also has encouraged me, when that was needed.

Contents

1 Introduction

In this thesis we are interested in seeing the difference in the genotype-disease association from including prevalent cases at baseline in the case-cohort analysis. The motivation stems from the MORGAM study in which the DNA from all prevalent cases, alive at baseline, was genotyped. Since the case-cohort analysis is valid for incident cases that occur during follow-up, the information from the prevalent cases at baseline was not used in the MORGAM study. The question remains whether the bias that these prevalent cases may have introduced in the case-cohort estimate of the genotype-disease association would have been outweighed by a gain in efficiency from the using the additional information from case genotypes at baseline. This thesis aims at introducing methodology that can shed some light on this question.

The prevalent cases in this thesis are the events of myocardial infarction, MI, that have happened before the study baseline for individuals who are still alive at baseline (age 45). The incident cases are the MI cases that happen during study follow-up, here from baseline (age 45) to censoring (age 80 or death, whichever comes first).

I generate populations according to two models, where the risk that an individual experiences an MI depends on genotype of that individual. For Model 0 the age at death with or without MI does not dependend on genotype while Model 1 assumes age at death after MI to depend on genotype and thus to induce selection for prevalent cases that are alive at baseline.

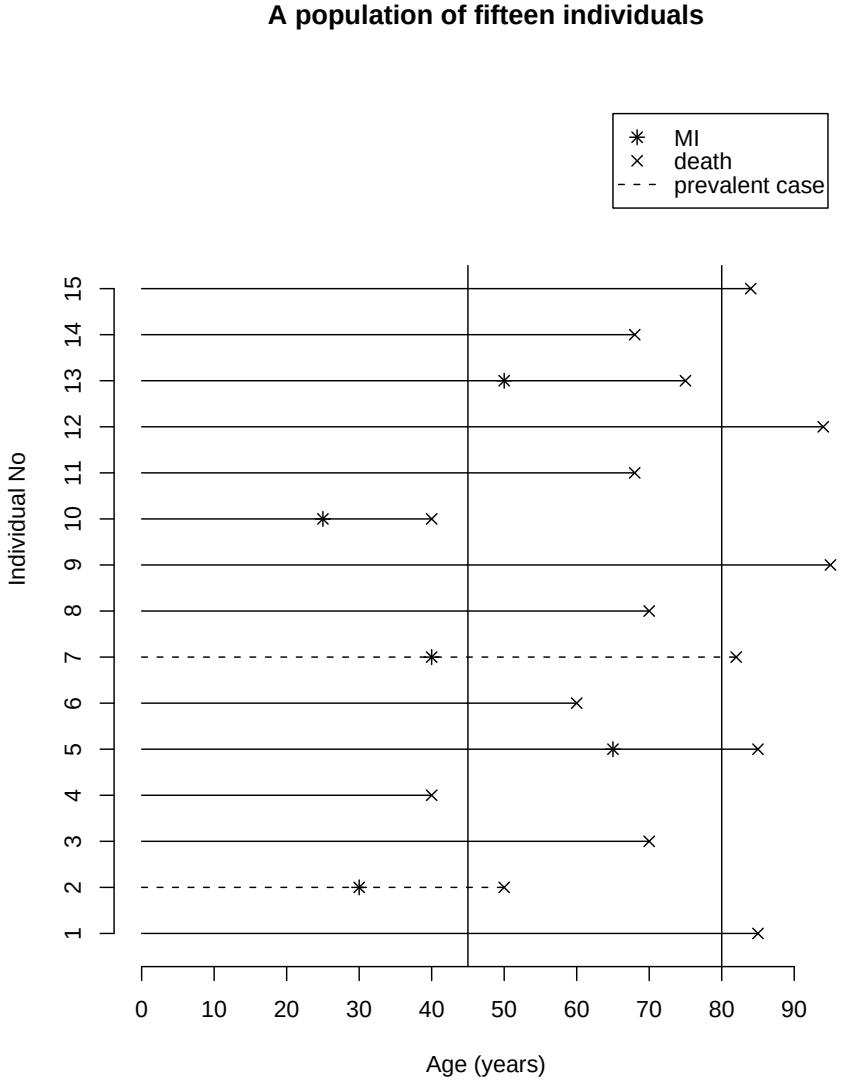
The populations I generate consist of 20 000 individuals. Figure 1 shows the structure for fifteen of these. We know when an individual dies, marked with x, and we know if and when an MI occurred. An MI is marked with *.

To study the properties of the models I simulate 1 000 replicates of each population. For all analyses I use the Cox proportional hazards regression model described in Section 2.1. However, for convenience I generate the data using a log linear Weibull model, described in Section 2.2, utilizing the fact that regression estimates and their standard errors coincide for the Cox regression model and the Weibull log-linear model. The traditional case-cohort analysis is introduced in Section 2.3 together with a description of how to include prevalent cases at baseline. Moreover, I use the illness-death model framework in my simulations to induce death rates after MI that depend on genotype through the

age at which the MI occurred. The illness-death model is presented in Section 2.4.

The detailed simulations are presented in chapter 3, with results and discussion in chapters 4 and 5. The data and R-code are included in Appendices.

Figure 1: A small population
*Here we follow the fifteen individuals from birth until death. We can see if they have an MI. The MI is marked with a *.*



1.1 The MORGAM project

The MORGAM project is a study on determinants for cardiovascular disease. The name MORGAM stands for **MO**NICA, **R**isk, **G**enetics, **A**rchiving and **M**onography. MONICA was a WHO (The World Health Organization) project about the risk factors for cardiovascular diseases. The name MONICA stands for Multinational MONItoring of trends and determinants in CArdiovascular disease. The MORGAM project is an extension of the MONICA project and includes genetic factors and also includes other cohorts than the ones in the MONICA project, as well as extensive biomaterial collection. There are mainly European countries in the MORGAM project, with the populations from different geographic areas. Australia, Denmark, Finland, France, Italy, Lithuania, Northern Ireland, Poland, Russia, Scotland, Sweden and Wales are areas that contribute cohorts. A local ethics committee has approved the study and participants have given informed consent. The samples and data are all processed anonymously.

DNA is taken from blood in a random sample of the full cohort, from all deaths and cardiovascular cases. Information of the DNA is collected for both the incident cases and the prevalent cases.

One purpose of the MORGAM project is to find the association between genetic variants and coronary heart disease and stroke. These diseases are called complex, multifactorial diseases because they are not caused by a single genetic defect, but by joint action from many genetic and environmental factors.

Information collected at the baseline, when individuals entered the project, included for example smoking, alcohol use, socioeconomic indicators, history of coronary heart disease, stroke, diabetes, family history of myocardial infarction and stroke. Anthropometric measurements, blood pressure, cholesterol, triglycerides, fibrinogen and SNP[1] genotype were also measured at the baseline. Triglycerides are fatty acids, where fat exists, and fibrinogen make clots of blood.

Different MORGAM centers, have used different follow-up methods on death. In some centers information on death was retrieved from the national death register and in other centers by periodic follow-up by letters or health care systems. The follow-up on the coronary and stroke events were retrieved from

[1] explanation in Section 1.2.1 on the following page

the MONICA register, hospitals discharge register, clinical event questionnaire and regional health information system. [10]

1.1.1 Myocardial infarction - Cardiovascular disease

A heart attack or an acute myocardial infarction, MI, occurs when the heart gets less blood supply than it should. The heart tissue is damaged and could die because of oxygen shortage, ischemia.

The disease is a common cause of death all over the world, for both men and women. The risk of an MI is higher for men at age 40 or older and women age 50 or older compared to younger men and women. There is a higher risk of an MI if the individual has had vascular disease. Other things that increase the risk of an MI are previous heart attack or stroke, abnormal heart rhythms or fainting, smoking, extreme alcohol consumption, abuse of several illegal drugs, high triglyceride levels, high LDL or low HDL (low- or high density lipoprotein), diabetes, high blood pressure, obesity and stress.

The name, myocardial infarction, comes from the heart muscle, *myocardium*, and tissue death due to oxygen starvation, *infarction*. Sometimes the name "heart attack" is used to describe sudden cardiac death and that might be an MI, but could also be some other type of heart failure. [2]

1.2 Genetic concepts and terminology

For the mathematician reader, with little background in biology or genetics, the central concepts in genetics used in this thesis will be explained.

1.2.1 Genetic terms

The human genome consists of *chromosomes,* which are DNA molecules. The *DNA* molecule, **d**eoxyribo**n**ucleic **a**cid, consists of two poly-nucleotide chains which are kept together by hydrogen bonds. The *genotype* is the specific gene set for an individual. [15]

The DNA molecule is built up of the *nucleotide bases* adenine, A, guanine, G, cytosine, C, and thymine, T. The bases A and G, as well as C and T respectively are complementary on a strand. These four bases occur linearly to form a DNA

sequence. A triplet of the bases is a *codon* and this is coding for an amino acid. Linearly arranged amino acids form specific proteins. [16]

A *gene* is a part of the DNA sequence that is coding for a polypeptide. Many polypeptides form a protein. Variants of a gene, in a specific chromosomal locus, on one chromosomal strand is called an *allele*. You need two alleles to form a gene. The place where a gene is located on a chromosome is called *locus*, (pl. loci). The genotype is *heterozygous* if the alleles differ, and *homozygous* if they are similar.

A *phenotype* is a property that is observable and may be correlated with the genotype. Here, we focus on cardiovascular disease phenotype such as MI, and consider association with genotypes. *Polymorphism* is the occurrence of more than one allele at a locus (form is morph in Greek) in a population. [15] A variation in the population involving a single nucleotide, that DNA is called SNP, Single Nucleotide Polymorphism. SNPs typically involves two alleles. Such variations could affect how individuals develop diseases. [3]

2 Survival and event history analysis

Event history analysis is used when one is interested in the occurrence of events over time. An event could be medical, such as death, myocardial infarction (MI) or cancer diagnosis, or non-medical, such as electric failure, divorce or birth of a child. In this thesis, MI constitutes the event of interest.

Event history analysis models is used to get information of the cause of the event in terms of risk factors. *Survival analysis* is describing the event process for a group of individuals by survival curves and hazard rates, and uses regression models to analyze the dependence on covariates. *Covariates* are the measured variables, that the event could be caused by or they could increase or decrease the risk for an event. In a survival model for MI one could, for example, include the covariates sex, age, weight, fitness and genotype. The result from event history analysis could be used to see how the covariates affect the event, MI. [12]

A survival function and a hazard function can describe *survival data*, data on the times for individuals until an event happens. Some of the survival times may be censored. *Censoring* occurs when an individual is lost to follow-up or

the individual does not reach the specific event for other reasons during the follow-up. The causes could be that the individual died for another reason than the event, that the data collectors could not come in contact with the individual or that the event had not happened when the study ended. The calendar time period when an individual is in the study is called the *study time*. The time from when the individual starts to participate in the study until the event happens is called *survival time*. For censored individuals the survival time is only partly observed. [9]

The *survival function*, $S(t)$, is the probability that an individual has not experienced the event by time t. We write

$$S(t) = P(T \geq t)$$

where the random variable T is survival time. The random variable T has the function $F(t) = P(T < t) = \int_0^t f(s)\,ds$, where $f(t)$ is the underlying probability *density function* of T. We also write the survival function as

$$S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t)$$

The *hazard function* or hazard rate, $\alpha(t)$, is the instantaneous probability density that an individual has the event at the time t if it is known that the individual "survived" (did not have the event) before that time. We write

$$\alpha(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \tag{1}$$

where T is a the survival time. [9, 12]

The survival function and the hazard function are connected through

$$A(t) = -\ln S(t)$$

where $A(t) = \int_0^t \alpha(s)\,ds$, is the *cumulative hazard*. To show this, we start with the definition of the hazard function in (1) and rewrite the numerator of (1) as

$$P(t \leq T < t + \delta t \mid T \geq t) = \frac{P((t \leq T < t + \delta t) \cap (T \geq t))}{P(T \geq t)} \tag{2}$$

According to the rule of conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

the nominator in (2) is simplified

$$P((t \leq T < t + \delta t) \cap (T \geq t)) = P(t \leq T < t + \delta t)$$

because $T \geq t$ does not provide any new information. We rewrite the numerator of (2) as

$$P(t \leq T < t + \delta t) = P(T < t + \delta t) - P(t > T) = F(t + \delta t) - F(t)$$

so equation (2) could be written as

$$\frac{P(t \leq T < t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

From this we get the hazard function

$$\alpha(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} = \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}$$

where the last equality sign comes from identifying the derivative of $F(t)$, which is $f(t)$. Now we have $\alpha(t) = \frac{f(t)}{S(t)}$. From $S(t) = 1 - F(t)$ we get $S'(t) = -f(t)$, so we have $\alpha(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}(\ln S(t))$ which by integrating gives us

$$A(t) = -\ln S(t)$$

□

Now it is showed how the survival function and the hazard function are connected. [9]

2.1 Proportional hazard model

The proportional hazard model or the *Cox regression model* is the basic model for survival data. The Cox regression model is semi-parametric because the baseline hazard is non-parametric and the relative risk function is parametric.

In the general proportional hazard model, the hazards of an event at a particular time depends on the values $x_1, x_2, \ldots, x_p$. These values are the covariates, recorded at the baseline. Each individual has his/her specific baseline. To handle a covariate that changes over time is more difficult and will not be discussed further here.

The hazard function of the i^{th} individual is

$$\alpha_i(t) = \psi(\mathbf{x}_i)\,\alpha_0(t)$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ are the p covariates for individual i and $\alpha_0(t)$ is the baseline hazard. The baseline hazard is a hazard function for an individual for whom all the covariates are zero. The relative hazard can not be zero, so it can be written as $\psi(\mathbf{x}_i) = e^{\eta_i}$ where η_i is a linear combination of all the covariates for individual i

$$\eta_i = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \sum_{j=1}^{p} \beta_j x_j$$

with β as the coefficients of the covariates. We write the general proportional hazards model as

$$\alpha_i(t) = e^{\eta_i}\alpha_0(t)$$

and we could rewrite that as

$$\ln\left(\frac{\alpha_i(t)}{\alpha_0(t)}\right) = \sum_{j=1}^{p} \beta_j x_{ji} = \beta^T \mathbf{x}_i$$

where $j = 1, \ldots, p$ denotes covariates. No assumptions have been made about the form of the baseline hazard function $\alpha_0(t)$. [9]

2.1.1 Partial likelihood

The hazard rate $\alpha(t \mid \mathbf{x}_i)$, with $\mathbf{x}_i$ the covariates for individual i, can be written as

$$\alpha(t \mid \mathbf{x}_i) = \alpha_o(t)\, r(\beta, \mathbf{x}_i(t)) \tag{3}$$

where $r(\beta, \mathbf{x}_i(t))$ is the relative risk function with $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ that describes the effect of the covariates, and $\alpha_0(t)$ is the baseline hazard. The $r(\beta, \mathbf{x}_i(t))$ is normalized, $r(\beta, \mathbf{0}) = 1$. For the Cox regression model the rel-

ative risk $r(\beta, \mathbf{x}_i(t)) = e^{\beta^{\mathbf{T}}\mathbf{x_i(t)}}$. Because the Cox regression model is semi-parametric, the partial likelihood turned out to be an efficient tool for estimating $\beta_1, \ldots, \beta_p$. It can be treated much as an ordinary likelihood. The partial likelihood has the form

$$L(\beta) = \prod_{T_j} \frac{Y_{i_j}(T_j)\, r\left(\beta, \mathbf{x}_{i_j}(T_j)\right)}{\sum_{l=1}^{n} Y_l(T_j)\, r(\beta, \mathbf{x}_l(T_j))} \tag{4}$$

where $Y_i(t)$ is an at-risk-indicator for individual i at time t, i_j is the index of the individual who experience the event at time T_j, and $r(\beta, \mathbf{x}_i(t))$ is the relative risk function. The at-risk-indicator, $Y_i(t)$, is

$$Y_i(t) = \begin{cases} 1 & \text{if at risk} \\ 0 & \text{if not at risk} \end{cases}$$

The partial likelihood is used to obtain the estimated β, by maximizing the function (4).

To derive the partial likelihood in formula (4) start with formula (3) and use $\lambda_i(t) = Y_i(t)\,\alpha(t \mid \mathbf{x}_i(t))$. From this

$$\begin{aligned} \lambda_i(t) &= Y_i(t)\,\alpha(t \mid \mathbf{x}_i(t)) = \\ &= Y_i(t)\,\alpha_o(t)\, r(\beta, \mathbf{x}_i(t)) \end{aligned}$$

The sum of all λ's is

$$\begin{aligned} \lambda_{\bullet}(t) &= \sum_{l=1}^{n} \lambda_l(t) = \\ &= \sum_{l=1}^{n} Y_l(t)\,\alpha_0(t)\, r(\beta, \mathbf{x}_i(t)) \end{aligned}$$

With

$$\begin{aligned} \pi(i \mid t) &= \frac{\lambda_i(t)}{\lambda_{\bullet}(t)} = \\ &= \frac{Y_i(t)\,\alpha_0(t)\, r(\beta, \mathbf{x}_i(t))}{\sum_{l=1}^{n} Y_l(t)\,\alpha_0(t)\, r(\beta, \mathbf{x}_l(t))} = \\ &= \frac{Y_i(t)\, r(\beta, \mathbf{x}_i(t))}{\sum_{l=1}^{n} Y_l(t)\, r(\beta, \mathbf{x}_l(t))} \end{aligned} \tag{5}$$

we get $\lambda_i(t) = \lambda_\bullet(t)\,\pi(i \mid t)$.

This, $\pi(i \mid t)$, is the conditional probability of observing an event for individual i at time t, given the past and given that an event is observed at that time. To obtain the partial likelihood for β, we take the product of all the conditional probabilities in equation (5) over all observed event times. Times when events are observed, $T_1 < T_2 < \ldots$. From this we have the partial likelihood function as in formula (4).

If we write the risk set at time T_j as $\mathcal{R}_j = \{l \mid Y_l(T_j) = 1\}$ the partial likelihood function from formula (4) can be rewritten as **[12]**

$$L(\beta) = \prod_{T_j} \frac{r\left(\beta, \mathbf{x}_{i_j}(T_j)\right)}{\sum_{l \in \mathcal{R}_j} r\left(\beta, \mathbf{x}_l(T_j)\right)} \tag{6}$$

2.2 Log-linear model

The data for simulations, in Chapter 3, is more conveniently generated with a log-linear model, than with the proportional hazard model. In certain situations, that we use, the two models are equivalent.

In a log-linear model the covariate directly expands or contracts the time to the event. The log-linear model can be written as

$$\ln t_i = \alpha + \beta^T \mathbf{x}_i + \sigma\epsilon_i \tag{7}$$

where t_i is the age or time for individual i. The $\mathbf{x}_i$ is a vector with the covariates for individual i, and the vector β are the coefficients to the covariates. The covariates can be genotype, age, sex, fitness etc, but in this thesis we have the covariates genotype and age. We will have only one covariate in each formula so β will be a constant. Therefore the ^{T}, that denotes a transpose, will be omitted.

The ϵ_i is extreme value distributed. The t_i is Weibull distributed with the two parameters, shape $\frac{1}{\sigma}$ and scale $e^{\alpha+\beta^T\mathbf{x}_i}$, according to the following derivation. [9]

The shape parameter $\frac{1}{\sigma}$ describes the form for the distribution. With $\sigma = 1$ we get an exponential distribution. If $\frac{1}{\sigma} = 3 - 3.5$ we get an approximately normal distribution. [4]

We now show of that t_i is Weibull distributed, when we know that ϵ_i is extreme value distributed. We want to show what distribution t_i has in formula

$$\ln t_i = \alpha + \beta^T x_i + \sigma\epsilon_i$$

with ϵ_i extreme value distributed.

Starting with the probability density function $f(\epsilon) = e^{\epsilon - e^{\epsilon}}$ for the extreme value distributed ϵ, where $-\infty < \epsilon < \infty$. Then we make a transformation from ϵ to t,

$$t_i = e^{\alpha + \beta^T x_i + \sigma\epsilon_i}$$

with $0 < t < \infty$. [9]

We will use that all probability density functions

$$\int_a^b f(x)\,dx = 1 \tag{8}$$

where $a < x < b$, and to remember to calculate dx. [5]

We write

$$\begin{aligned} \epsilon_i &= \frac{1}{\sigma}\left(\ln t_i - \alpha - \beta \mathbf{x}_i\right) \\ d\epsilon_i &= \frac{1}{\sigma} \cdot \frac{1}{t_i} dt_i \end{aligned}$$

and that with equation (8) we can write

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} f(\epsilon)\,d\epsilon &= \int_{-\infty}^{\infty} e^{\epsilon - e^{\epsilon}} d\epsilon = \\ &= \int_0^{\infty} e^{\frac{1}{\sigma}(\ln t_i - \alpha - \beta \mathbf{x}_i) - e^{\frac{1}{\sigma}(\ln t_i - \alpha - \beta \mathbf{x}_i)}} \frac{1}{\sigma} \cdot \frac{1}{t_i} dt_i = \\ &= \int_0^{\infty} t_i^{\frac{1}{\sigma}} e^{-\frac{1}{\sigma}(\alpha + \beta x_i)} e^{-e^{\frac{1}{\sigma}(\ln t_i - \alpha - \beta \mathbf{x}_i)}} \frac{1}{\sigma} \cdot \frac{1}{t_i} dt_i = \\ &= \left[a = e^{\alpha + \beta x_i} \text{ and } b = \frac{1}{\sigma}\right] = \\ &= \int_0^{\infty} t_i^{b-1} \cdot b \cdot \frac{1}{a^b} e^{-\left(\frac{t_i}{a}\right)^b} dt_i = \\ &= \int_0^{\infty} \frac{b}{a^b} t_i^{b-1} e^{-\left(\frac{t_i}{a}\right)^b} dt_i \end{aligned}$$

Comparing this result to the Weibull probability density function with two pa-

rameters, scale a and shape b,

$$f(x;a,b) = \int_0^\infty \frac{b}{a^b} x^{b-1} e^{-\left(\frac{x}{a}\right)^b} dx$$

we see that t_i is Weibull distributed with the two parameters, scale $e^{\alpha+\beta x_i}$ and shape $\frac{1}{\sigma}$.

□

Another log-linear model can be written

$$\ln t_i = \frac{1}{k}\left(-\ln \lambda_2 - \beta_2^T \mathbf{x}_i + \epsilon_i\right) \tag{9}$$

where t_i is the age or time for individual i. This t_i is Weibull distributed with the two parameters, shape $\frac{1}{\sigma}$ and scale $e^{\frac{1}{k}\left(-\ln \lambda_2 - \beta_2^T \mathbf{x}_i\right)}$. The $\mathbf{x}_i$ is a vector with the covariates for individual i, and the vector β_2 times $\frac{1}{k}$ are the coefficients to the covariates. The ϵ_i is extreme value distributed.

The Cox regression analysis returns β_2. I use the model in equation (7). A comparison between the models in equation (7) on page 13, and in equation (9) we get

$$\begin{aligned} k &= \frac{1}{\sigma} \\ \beta_2 &= -\frac{\beta}{\sigma} \\ \lambda_2 &= e^{-\frac{\alpha}{\sigma}} \end{aligned}$$

Time for individual i, t_i, is given by exponating equation (7),

$$t_i = e^{\alpha+\beta^T \mathbf{x}_i + \sigma\epsilon_i} \tag{10}$$

The time is almost always age in this thesis. [9]

2.2.1 How to generate ϵ_i

The formula for ϵ_i is

$$\epsilon = \ln(-\ln(1-p))$$

where p is a probability between 0 and 1, uniformly distributed. This formula is obtained from the probability density function for the extreme value distribution

$$f(\epsilon) = e^{h(\epsilon)} \text{ and } h(\epsilon) = \epsilon - e^{\epsilon} \text{ with } -\infty < \epsilon < \infty$$

The transformation $\xi = e^{\epsilon}$ helps to obtain ϵ_i. We obtain the probability density function $g(\xi) = e^{-\xi}$ with $0 < \xi < \infty$. When we integrate this we get

$$\begin{aligned} p &= \int_0^{\xi} g(u)\, du = \int_0^{\xi} e^{-u} du = \\ &= -e^{-u}|_{u=0}^{u=\xi} = -e^{-\xi} - \left(-e^{-0}\right) = \\ &= 1 - e^{-\xi} \end{aligned}$$

Here follows a derivation that p is uniformly distributed.

Assume that the variable $t \in [A, B]$ with a density function $f(t)$.

We know, from (8) on page 14, that

$$\int_A^B f(s)\, ds = 1 \tag{11}$$

We define

$$p(t) = P(T < t) = \int_A^t f(s)\, ds$$

We want to change the variable from s to p, $s(p)$ and $\frac{dp}{ds} = f(s)$ from the definition above, so $dp = f(s)\, ds$. Now we have

$$1 = \int_{p(A)}^{p(B)} dp = \int_{p(A)}^{p(B)} 1 dp$$

where 1 is a constant density function for p. We have that $p(B) = 1$, because this is the integral over the whole set from A to B. The lower bound, $p(A) = 0$, because it is the integral from A to A. The probability density function for an individual with a uniform distribution [6] between a and b is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Here we can see that the probability density function for p is

$$1 = \frac{1}{p(B) - p(A)} = \frac{1}{1-0}$$

so p is uniformly distributed between 0 and 1, $p \sim U(0,1)$.

□

To first obtain ξ we use

$$\begin{aligned} p &= 1 - e^{-\xi} \\ 1-p &= e^{-\xi} = \frac{1}{e^{\xi}} \\ e^{\xi} &= \frac{1}{1-p} \\ \xi &= \ln\left(\frac{1}{1-p}\right) = -\ln(1-p) \end{aligned}$$

Out of this we obtain ϵ by formula (12) where we generate p randomly between 0 and 1 from a uniform distribution.

$$\begin{aligned} \epsilon &= \ln \xi = \\ &= \ln(-\ln(1-p)) \end{aligned} \tag{12}$$

2.2.2 Expected value of t_i

To determine the parameters, α and β, we need to use the expected value of equation (7) on page 13

$$\begin{aligned} E(t_i) &= E\left(e^{\alpha+\beta'\mathbf{x}_i+\sigma\epsilon_i}\right) = \\ &= e^{\alpha+\beta'\mathbf{x}_i} \cdot E(e^{\sigma\epsilon_i}) \end{aligned}$$

The expected value of $e^{\sigma\epsilon_i}$ is $\Gamma(\sigma+1)$, where Γ is the gamma function, according to the following equations

$$
\begin{aligned}
E\left(e^{\sigma\epsilon}\right) &= \int_{-\infty}^{\infty} e^{\sigma\epsilon}\cdot e^{\epsilon-e^{\epsilon}}d\epsilon = \\
&= \left[E\left(g\left(\epsilon\right)\right)=\int g\left(\epsilon\right)f\left(\epsilon\right)d\epsilon\right] = \\
&= \int_{-\infty}^{\infty} e^{(\sigma+1)\epsilon}\cdot e^{-e^{\epsilon}}d\epsilon = \\
&= \left[\xi=e^{\epsilon}, d\xi=e^{-\epsilon}d\epsilon \Longleftrightarrow d\epsilon=\frac{1}{\xi}d\xi, -\infty<\epsilon<\infty, 0<\xi<\infty\right] = \\
&= \int_{0}^{\infty} \xi^{\sigma+1}e^{-\xi}\frac{1}{\xi}d\xi = \\
&= \int_{0}^{\infty} \xi^{\sigma}e^{-\xi}d\xi
\end{aligned}
$$

To solve the integral $\int_0^\infty \xi^\sigma e^{-\xi}d\xi$ we use the formula

$$\int_0^\infty x^n e^{-ax}dx = \frac{1}{a^{n+1}}\Gamma\left(n+1\right)$$

where Γ is the gamma function. [8]

The integral is

$$
\begin{aligned}
\int_0^\infty \xi^\sigma e^{-\xi}d\xi &= \frac{1}{1^{\sigma+1}}\Gamma\left(\sigma+1\right) \\
&= \Gamma\left(\sigma+1\right)
\end{aligned}
$$

We get

$$E\left(e^{\sigma\epsilon_i}\right) = \Gamma\left(\sigma+1\right) \tag{13}$$

To calculate Γ, the `gamma` function in the software R will be used, and then σ can be any positive real number.

□

2.3 The Case-cohort study

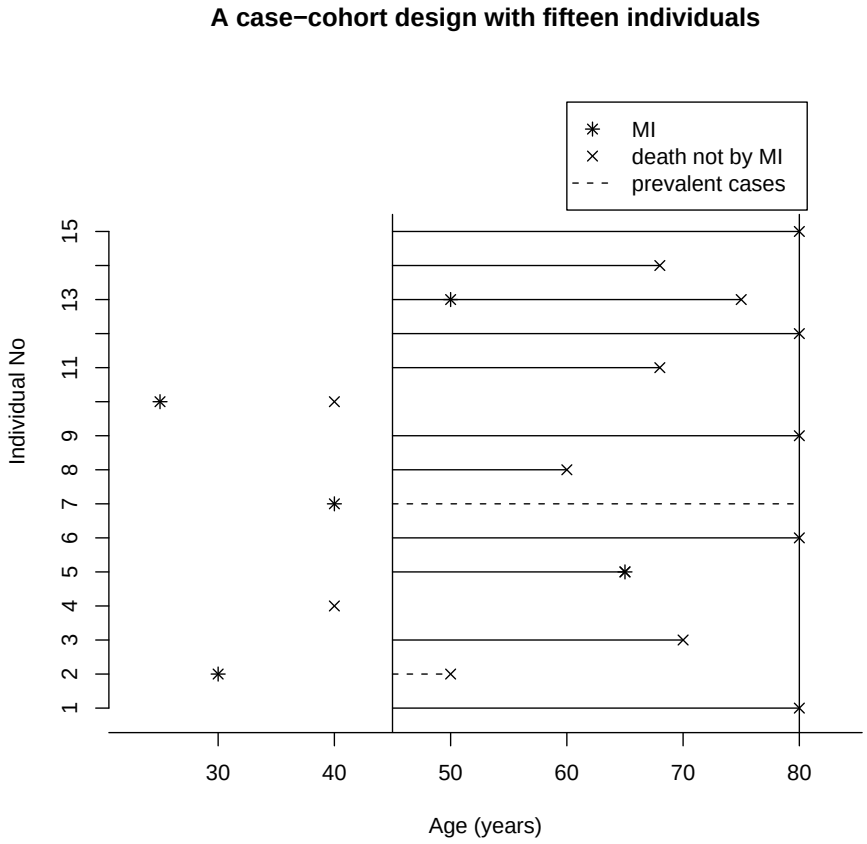
The case-cohort study design is a method for studying time-to-event-data without needing to collect covariate information on all individuals. Here, it is substantially cheaper to collect DNA samples on few individuals. It is only needed

to collect DNA for all individuals who experienced the event and for a subcohort of all individuals in the study. The latter *subcohort* is a randomly chosen sample from the whole population. It is important that the subcohort is chosen without looking at the covariates that we think contribute to the event, MI. The subcohort is a comparison group for all the MI cases in the cohort. In most of the case-cohort studies, information for the covariates is collected when the individual enters the study. For genetic studies DNA can be collected at any time during the study. Since DNA is stable over an individual's lifespan DNA can be collected at any time during the study. For MI cases DNA is collected at the time of diagnosis. To analyze the case-cohort samples there are several methods, analogous to methods for the full cohort data. [13] Here we use the partial likelihood described in Section 2.3.1.

In Figure 2 we follow fifteen hypothetical individuals from when they enter the study to an MI or a death. We can also see if they had an MI or not. The death is marked with an x and an MI is marked with a $*$. The prevalent cases, individuals who had an MI before baseline, is at baseline marked with a ■.

Figure 2: Fifteen individuals in the study

We follow the same individuals as in Figure 1 from the time they enter the study at the baseline, age 45. Two individuals do not reach the age of 45, so in the study we do not even know that they existed. We follow the remaining thirteen individuals until they get an MI, die or are censored at age 80.



2.3.1 Partial likelihood function for case-cohort design

The partial likelihood for case-cohort design is obtained from formula (6) with different sets for the sums in the denominator,

$$\widetilde{L}(\beta) = \prod_{T_j} \frac{r\left(\beta, \mathbf{x}_{i_j}(T_j)\right)}{\sum_{l \in \widetilde{\mathcal{R}}_j(t)} r\left(\beta, \mathbf{x}_l(T_j)\right)}$$

where $\widetilde{\mathcal{R}_j}(t)$ is the case-cohort set and consist of the chosen subcohort and the MI cases outside the subcohort, $\widetilde{\mathcal{R}_j}(t) = \widetilde{C}(t) \cup \{i_j\}$. The $\widetilde{C}(t)$ is the subcohort at time t, where individuals who had an MI are removed after the MI has occurred. The $\{i_j\}$ is the set of the MI case that occurs at time T_j. [14]

2.3.2 Including prevalent cases

Figure 2 presents prevalent cases that have occurred before baseline for individuals that are alive at baseline. Each of these prevalent cases contributes a term to the partial likelihood with their genotype in the numerator and the denominator summed over the subcohort at baseline enriched with the prevalent cases.

2.4 The illness-death model

We use the illness-death model to introduce death after MI that depends on age at MI and thus, indirectly, on the genotype. This should introduce selection and bias in using prevalent cases at baseline in the case-cohort analysis. The illness-death model has a Markov property.

A *Markov chain* is a stochastic process with discrete states and discrete time, $\{X_1, X_2, \ldots\}$ where X_n is a discrete stochastic variable, and fulfills

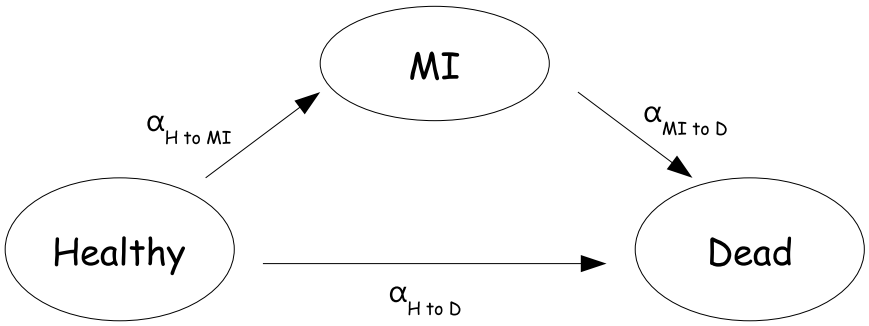
$$P(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{i-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i_n)$$

where $j, i, i_{n-1}, \ldots$ are different states. There are Markov chains, called Markov processes, that are time continuous, but we will not use them in this thesis. This equation means that the future is not dependent on the past, it only depends on the present state. [11]

In this thesis, the illness-death model has three states, as in Figure 3, individuals that are healthy (no MI or death) in state "Healthy", individuals who have had an MI (and not yet died) in state "MI" and individuals who are dead (no matter an MI or not) in state "Dead".

Figure 3: Illness-death model

In this illness-death model there are three states. The transition intensities between the states are marked with α.



In this figure the α's are transition intensities, the instantaneous risk of moving from state "Healthy" to state "MI" is denoted $\alpha_{\mathrm{H\,to\,MI}}$ and so on.

The probability that an individual is in state "a" at time t_1 and is in state "b" at a later time t_2 is written $P_{ab}(t_1, t_2)$. This probability can be written as

$$P_{ab}(t_1, t_2) = P(X(t_2) = b \mid X(t_1) = a)$$

where a, b are different states in a Markov chain, and $t_1 < t_2$ as said above. We have the transition intensity

$$\alpha_{ab}(t) = \lim_{\Delta t \to 0} P(X(t + dt) = b \mid X(t-) = a)$$

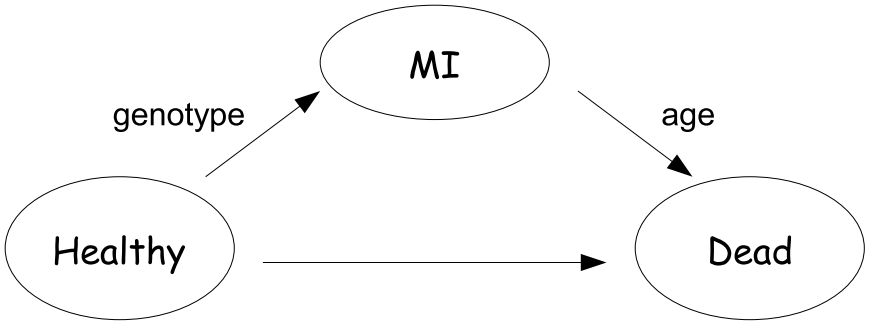
where a and b are two states and t is the time. [12]

3 The simulation study

I simulate a population based on the illness-death model and study the effect of including prevalent cases at baseline when evaluating the MI, in the case-cohort design. I use the software R [1]. There are three states in this illness-death model, see Figure 4.

In my models the genotype may affect transition from state "Healthy" to state "MI", $\alpha_{\text{H to MI}}$, and age of the MI may affect the transition from state "MI" to state "Dead", $\alpha_{\text{MI to D}}$. The risk of dying for an individual, who has not had an MI, is smaller than the risk of dying for an individual, who had an MI. That is, the risk of transition from state "Healthy" to state "Dead" is smaller than the risk of transition from state "MI" to state "Dead". The genotype is not assumed to directly affect the risk of transition from state "MI" to state "Dead".

Figure 4: Illness-death model for MI



3.1 Generating data

I generate a population of 20 000 individuals from birth with information on age of death for each individual, an indicator if the individual has had an MI or not and the age when the individual had an MI. My data also consists of information on the genotype and an indicator if the individual is in the subcohort. The subcohort is every 10^{th} individual in the population. An example of data can be seen in Table 8 on page 38.

First I generate an age of natural death for each individual using the log-linear model in formula (7), on page 13, with the procedure to generate ϵ_i described in Section 2.2.1 on page 15. Natural death means other causes of death than an MI. All times are in years.

Then I generate an age when each individual gets an MI, and time until death after their MI. I assume that an individual can get at most one MI. Now each individual have age for two deaths, the age of natural death, and the age of

death after an MI. The actual age of death will be the age of whatever kind of death that occurs first for each individual.

There are two models for the transition from state "Healthy" to state "Dead" via state "MI". These two models will be described in detail below. The natural death is the same for both models.

3.1.1 Genotypes

The genotype is the covariate of interest in this thesis. As mentioned in Section 1.1, on page 6, the MI is a multifactorial disease. Here we assume a so called "candidate SNP" scenario, and study one genotype at a time. In my two models the age when the individual gets an MI is assumed to depend on the genotype of the individual. The individuals inherit their alleles from their parents, one from each parent. To generate the genotype of an individual, first I simulate which of the alleles are inherited.

For simplicity, it is assumed that the parents are heterozygous, that their genotype is Aa, so they have one allele of each type. The value 0 represents allele a and 1 represents allele A. The allele inherited from each parent is either 0 or 1, binomial distributed with probability p. The chance of inheriting either allele is equal, so the probability is $p = 0.5$. To get the genotype for the offspring, sum the values for the alleles from the parents. The sum for the offspring is 0, 1 or 2 which represents genotype aa, Aa respective AA. I assume that the risk of an MI is greatest for genotype AA and least for the genotype aa. Therefore the β in the log-linear model, in equation (7) on page 13, will be negative.

3.1.2 Transition from state "Healthy" to state "Death"

The natural death is generated in the same way for Model 0 and 1, described below. To generate how natural death depends on age, transition from state "Healthy" to state "Dead", I use the log-linear model without any covariates. The age of natural death is denoted $t_i^{(death)}$ for individual i, and calculated by

$$t_i^{(death)} = e^{\alpha+\sigma\epsilon_i} \tag{14}$$

where $t_i^{(death)}$ is a random variable that follows a Weibull distribution with parameter scale e^{α}, shape $\frac{1}{\sigma}$ since ϵ_i an extreme value distribution.

To get a realistic value of the parameter α, I choose $\sigma = \frac{1}{9}$ and the mean age of natural death to be $\overline{t_i^{(death)}} = 78$ years, which is close to the average length of life. Then we use the formula

$$\begin{aligned} \overline{t_i^{(death)}} &= E\left(e^{\alpha+\sigma\epsilon_i}\right) = \\ &= e^{\alpha} \cdot E\left(e^{\sigma\epsilon_i}\right) = \\ &= e^{\alpha}\Gamma\left(\sigma+1\right) \end{aligned} \tag{15}$$

because of the result in equation (13).

Equation (15) gives

$$\alpha = \ln\left(\frac{\overline{t_i^{(death)}}}{\Gamma\left(\sigma+1\right)}\right)$$

To get ϵ_i, I use formula (12), on page 17, with p random uniformly distributed between 0 and 1. Now we have all the parameters needed to calculate the distribution of natural death from formula (14).

3.1.3 Transition from state "Healthy" to state "MI"

To generate the age when an individual has an MI, I use the log-linear model. The risk of getting an MI is set to depend on the genotype. Two age groups are distinguished depending on age when the MI occurs. The first group consists of those individuals who had an MI before the age of 45, "MI age $<$ 45". The second group consists of those who had an MI at age 45 or later, "MI age $\geq$ 45". The transition from state "Healthy" to "Death" in Model 1 is different from Model 0, described below. For Model 0 the relative risk of getting an MI does not change for the two age groups. But for the other two models the relative risk of getting an MI depending on genotype is higher before age 45 than after.

3.1.4 Transition from state "MI" to state "Death"

For Model 0 the risk of dying is the same regardless of age group when MI occurred. In Model 1, the risk of dying after an MI is higher for an individual who experienced an MI at young age, before age 45.

3.1.5 Data structure

When we know the age of possible death, a comparison for each individual is made between the age of natural death and the age of death by MI. The age of real death is the age of the first possible death that happens to the individual. All individuals are censored at age 80. For individuals who have not had an MI before age 80, the age of MI is not available, NA. The age of death is known for all the individuals, it is known if they had an MI or not, the age of MI, and if they were censored or not.

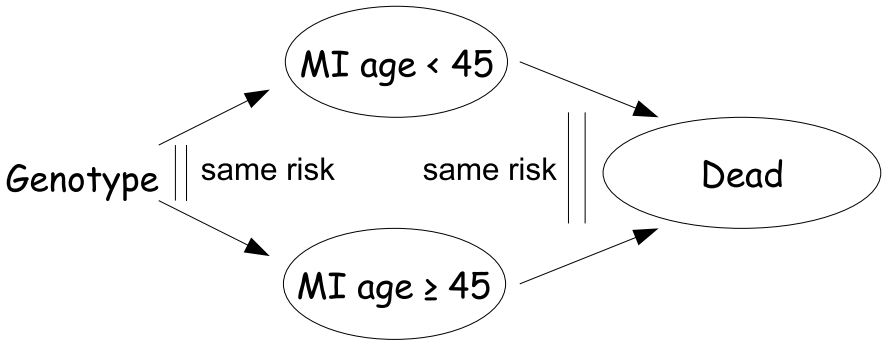
In our data we will have for each individual, age of death, the genotype, MI indicator and age of MI if the individual has had an MI. There are three different indicator values, the indicator value 0 means that the individual has not had an MI, 1 means that the individual has had an MI and 2 means that the individual was censored. An example of the ten first individuals in Model 0 is in Table 8 in Section A on page 38.

3.1.6 Model 0

In this model, see Figure 5, the risk of getting an MI depends on genotype only. Further more, the relative risk of getting an MI depending on the genotype before age 45, is the same as the relative risk of getting an MI after age 45. This means that the parameter β is the same for the two age groups. The risk of dying given an MI is not here depending on age group of when MI occurred.

Figure 5: Model 0

In this model the relative risk of getting an MI does not change with age, $\beta_{(<45)} = \beta_{(\geq 45)}$. *The risk of dying is also the same regardless of age group.*



Transition from state "Healthy" to state "MI"

To generate a vector containing age when an individual has his/her MI I use a variant of formula (7) on page 13,

$$t_i^{(MI)} = e^{\alpha+\beta\cdot G_i+\sigma\epsilon_i} \tag{16}$$

where $t_i^{(MI)}$ is Weibull distributed with the two parameters scale and shape, $e^{\alpha+\beta\cdot G_i}$ respective $\frac{1}{\sigma}$. The G_i is the genotype of individual i (0 for aa, 1 for Aa and 2 for AA), ϵ_i is extreme value distributed. I adjust the parameters σ, β and α to get a reasonable age distribution when the MI occurs. I choose $\sigma = \frac{1}{8}$, and this value is used for all following σ's.

To get the values ϵ_i, use formula (12), on page 17, and generate the probability p with uniform distribution.

For genotype aa, the age when MI occurs is generated from

$$t_{aa_i}^{(MI)} = e^{\alpha+\beta\cdot 0+\sigma\epsilon_i}$$

with mean age when the MI occurs for this genotype, $\overline{t_{aa}} = 90$. We get α from

$$\alpha = \ln\left(\frac{\overline{t_{aa}}}{\Gamma(\sigma+1)}\right)$$

Now we have the parameters α and ϵ_i.

The distribution of age at MI for the other two genotypes, Aa and AA, determines the parameter β. I choose a mean age when the MI occurs for genotype Aa, $\overline{t_{Aa}} = 75$. and obtain the parameter β from

$$\begin{aligned} \beta &= \ln\left(\frac{\overline{t_{Aa}}}{\Gamma(\sigma+1)}\right) - \alpha = \\ &= \ln\left(\frac{\overline{t_{Aa}}}{\Gamma(\sigma+1)}\right) - \ln\left(\frac{\overline{t_{aa}}}{\Gamma(\sigma+1)}\right) = \\ &= \ln\left(\frac{\overline{t_{Aa}}}{\overline{t_{aa}}}\right) \end{aligned} \tag{17}$$

With all the parameters defined and the genotype vector generated, I use formula (16) to generate the age when MI occurs.

Transition from state "MI" to state "Dead"

To generate a vector with age of death via MI, I use a form of the formula (10), on page 15,

$$t_i^{(death|MI)} = e^{\alpha+\sigma\epsilon_i} \tag{18}$$

The parameter ϵ_i is generated, as before, with formula (12), on page 17, and the probability p, which is a uniform distribution.

To set the parameter α, I choose a mean time the individuals live after an MI, $\bar{t} = 10$. After setting this value in the formula below, we obtain the parameter α.

$$\begin{aligned} \bar{t} &= E\left(e^{\alpha+\sigma\epsilon_i}\right) \\ &= e^{\alpha}\Gamma(\sigma+1) \end{aligned}$$

I choose the value of σ to get a realistic age distribution, $\sigma = \frac{1}{8}$.

From this we obtain the parameter α as

$$\alpha = \ln\left(\frac{\bar{t}}{\Gamma(\sigma+1)}\right) \tag{19}$$

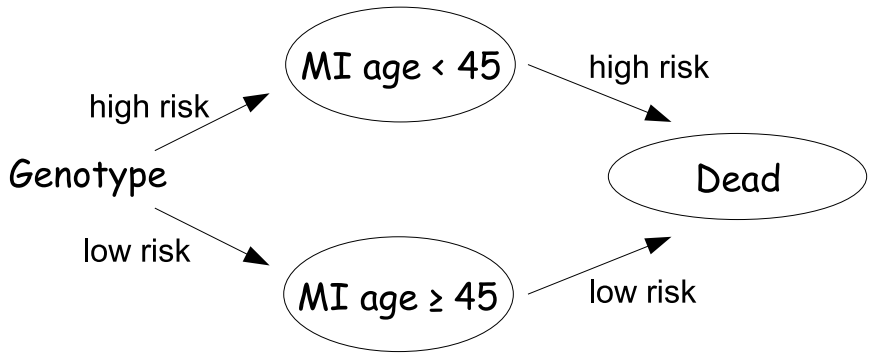
Now when we have all the parameters we use formula (18).

3.1.7 Model 1

In this model the risks are different, see Figure 6. The relative risk of getting an MI depending on the genotype before age 45, is higher than the relative risk of getting an MI after age 45. The risk of dying for an individual is greater in state "MI age < 45" than in state "MI age ≥ 45".

Figure 6: Model 1

All the risks are different. The relative risk of getting an MI is higher before age 45 than after. The risk of dying is different for the individuals in the two age groups.



Transition from state "Healthy" to state "MI"
To generate age when an MI occurs, we use formula (16), on page 27, as in Model 0. The difference between Model 0 and Model 1 is that the risk that depends on genotype in Model 0 is the same and in Model 1 is different. We generate a preliminary vector with age when MI occurs with formula

$$t_{prel_i}^{(MI)} = e^{\alpha+\beta\cdot G_i+\sigma\epsilon_i} \tag{20}$$

where $t_{prel_i}^{(MI)}$ is Weibull distributed with the two parameters scale and shape, $e^{\alpha+\beta\cdot G_i}$ respective $\frac{1}{\sigma}$. The G_i is the genotype of individual i (0 for aa, 1 for Aa and 2 for AA), ϵ_i is extreme value distributed. To obtain the values for α, β and ϵ_i I do as in Model 0, using formulas (19),(17) on page 28 and (12) on page 17, with the same values on σ, $\overline{t_{aa}}$ and $\overline{t_{Aa}}$.

To differentiate how the genotype effect depends on age, use formula (20) again for all values in the preliminary vector, $t_{prel_i}^{(MI)}$ which are greater than 45. The

age when the MI occurs is obtained using the following formula

$$\begin{aligned} t_i^{(MI)} &= I\left(t_{prel_i}^{(MI)} < 45\right) \cdot t_{prel_i}^{(MI)} + \\ &\quad + I\left(t_{prel_i}^{(MI)} \geq 45\right) \cdot \left(e^{\alpha+\beta_2 \cdot G_i + \sigma\epsilon_i^{\cdot}}\right) \end{aligned} \tag{21}$$

where $t_{prel_i}^{(MI)}$ is the preliminary age when individual i got an MI and $I(x) = 1$ if x is true and 0 otherwise. The parameter β_2 is here chosen to be the same as β, but it could have bee chosen to be another value than β. The $\epsilon_i^{\cdot}$ is regenerated, using the same procedure as above for ϵ_i.

Now we can calculate the age when the MI occurs using equation (21).

Transition from state "MI" to state "Dead" To get the distribution of age of death given an MI, that is the transition from state "Healthy" to state "Dead" via state "MI". To get the time from the MI until death I use formula

$$e^{\alpha+\beta \cdot I_i(MI \geq 45)+\sigma\epsilon_i}$$

So I simulate the age of death given an MI using

$$\begin{aligned} t_i^{(death|MI)} = & \quad t_i^{(MI)} + \\ & + \quad e^{\alpha+\beta \cdot I_i(MI \geq 45)+\sigma\epsilon_i} \end{aligned} \tag{22}$$

where the indicator

$$I_i\left(MI \geq 45\right) = \begin{cases} 0 & \text{if } t_i^{(MI)} < 45 \\ 1 & \text{if } t_i^{(MI)} \geq 45 \end{cases}$$

Now we want to obtain the parameters α, β and ϵ_i. I choose $\sigma = \frac{1}{8}$, as before. To obtain ϵ_i, I do as before, using formula (12), on page 17. To obtain α and β, I do as in Model 0, and obtain formulas (19) and (17), on page 28. I use the same value for with the same value on σ. But the risk for the two age groups are different, and time from MI until death is $\overline{t_{(<45)}} = 5$ and $\overline{t_{(\geq 45)}} = 10$.

We have the formula

$$\begin{aligned}\overline{t_{(<45)}} &= E\left(e^{\alpha+\beta\cdot 0+\sigma\epsilon_i}\right) = \\ &= e^{\alpha}E\left(e^{\sigma\epsilon_i}\right) = \\ &= e^{\alpha}\Gamma\left(\sigma+1\right)\end{aligned}$$

and

$$\begin{aligned}\overline{t_{(\geq 45)}} &= E\left(e^{\alpha+\beta\cdot 1+\sigma\epsilon_i}\right) = \\ &= e^{\alpha+\beta}E\left(e^{\sigma\epsilon_i}\right) = \\ &= e^{\alpha+\beta}\Gamma\left(\sigma+1\right)\end{aligned}$$

From these two formulas we obtain

$$\alpha = \ln\left(\frac{\overline{t_{(<45)}}}{\Gamma\left(\sigma+1\right)}\right)$$

and

$$\beta = \ln\left(\frac{\overline{t_{(\geq 45)}}}{\overline{t_{(<45)}}}\right)$$

Now, to get age of death given an MI, we use formula (22).

3.1.8 Input data

The input data for Model 0 and 1 are in Table 1, where we also can see the values of the times from the MI until death.

Table 1: Input data for Model 0 and 1

Input data for Model 0 and 1	
mean age of natural death	$\overline{t_i^{(death)}} = 78$
shape parameter for Weibull	$k = \frac{1}{\sigma} = 8$
mean age when MI occurs with genotype aa	$\overline{t_{aa}} = 90$
mean age when MI occurs with genotype Aa	$\overline{t_{Aa}} = 75$
censoring age	80

Time from the MI until death

In Model 0	**In Model 1**
$\overline{t_{MI \to Death}} = 10$	$\overline{t_{(<45)}} = 5$
	$\overline{t_{(\geq 45)}} = 10$

3.2 Analyzing data

To analyze the simulated data we use the package `survival` in the software R [1]. I compare the case-cohort analysis with and without the prevalent cases to the Cox regression model for the full cohort. I treat the mean over the 1 000 replicates of the Cox regression estimate for β as the true value to which I compare the case-cohort estimate of β and its standard error, seβ.

Note that I have introduced dependent censoring by death when I use the log linear model to generate the data according to the illness-death model. I therefore cannot expect to retrieve the input β for MI from the Cox model even when the full cohort is used. Instead I compare the two case-cohort scenarios to the generated Cox model for the full cohort.

3.2.1 Cox regression model analysis

The Cox regression model analysis is, in R code, called `coxph`. In the Cox analysis I use *all* MI cases in my population.

3.2.2 Case-cohort design analysis

The case-cohort design analysis is, in R code, called `cch`. In one of the case-cohort analyses I use only incident cases, and in the other case-cohort analysis I use both the incident cases and the prevalent cases. However both case-cohort

analyses only use the MI cases where individuals still are alive after age 45, baseline, since they are known. I also know the age of MI for the prevalent cases.

3.2.3 95 % coverage

To get the 95% coverage for both analyzing methods, we are testing if the true β is in the interval from the calculated β for Cox, $\hat{\beta}_{Cox}$, and for case-cohort design, $\hat{\beta}_{cch}$, plus-minus the standard errors for these. We test on level 95% that gives us the value 1.96. That is, we check how many times of the 1 000 the true β is inside the intervals,

$$\beta \in \left(\hat{\beta}_{Cox_i} \pm 1.96 \cdot se\hat{\beta}_{Cox_i}\right)$$

and

$$\beta \in \left(\hat{\beta}_{cch_i} \pm 1.96 \cdot se\hat{\beta}_{cch_i}\right)$$

where i is the number of the population, $i = 1, \ldots, 1\,000$. For both analyzing models, the real β is the mean of $\hat{\beta}_{Cox}$, where $\hat{\beta}_{Cox}$ are the 1 000 $\hat{\beta}$'s returned from the Cox regression analysis. The result is expressed in percent in Table 2 and 5.

3.2.4 Mean square error

To see the difference in precision and bias between the two case-cohort models, with and without the prevalent cases the *mean square error*, MSE, is used. Two mean square error is calculated, one for the case-cohort design without the prevalent cases, and on for the case-cohort design with the prevalent cases included. The formula for mean square error is

$$MSE\left(\hat{\beta}\right) = var\left(\hat{\beta}\right) + \left(bias\left(\hat{\beta}\right)\right)^2$$

where the bias and variance is calculated as below.

The bias is calculated with

$$bias\left(\hat{\beta}\right) = E\left(\hat{\beta}\right) - \beta$$

Table 3: Mean square error, variance and bias for Model 0

Analysis type	mean square error	variance	bias
Case-cohort without prevalent cases	0.0105	0.0105	-0.0086
Case-cohort with prevalent cases	0.0099	0.0099	-0.0032

where the $\hat{\beta}$ is the mean of the value from the case-cohort design and β is the true value of β and that is the mean of the β's from the Cox regression model.

The variance is

$$var\left(\hat{\beta}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\beta}_i - \text{mean}\left(\hat{\beta}\right)\right)^2$$

but is calculated with the function `var` in the software R. Here $\hat{\beta}_i$ is the value i from case-cohort design and $\hat{\beta}$ is a vector with all values from case-cohort design. [7]

4 Results

4.1 Results from Model 0

The results from Model 0 is presented in Table 2. There are the calculated "real" value on β which is set to be the same as the means of all 1 000 $\hat{\beta}$ for Cox regression analysis. We also have the mean of all 1 000 $\hat{\beta}$ for case-cohort analysis. In the table we also can see the mean of the standard errors, se $\hat{\beta}$, from the two analyzing methods, then the calculated standard deviation, sd $\hat{\beta}$, and the 95% coverage for both models.

Table 2: Results from Model 0
In the Cox regression model there are 20 000 individuals, but in the case-cohort analysis the subcohort is only 2 000 individuals. The true value of β =mean$\hat{\beta}_{Cox}$.

Analysis	mean $\hat{\beta}$	mean se$\hat{\beta}$	sd$\hat{\beta}$	95% coverage
Cox regression model	0.2734	0.0236	0.0229	0.963
Case-cohort	0.2648	0.1064	0.1023	0.954
Case-cohort with prevalent	0.2702	0.1038	0.0996	0.955

We could also be interested in knowing the percentage of individuals who are

healthy until death, who have an MI before they die or who are censored because they still live and are older than the censoring age, 80 years. These results we can see in Table 4.

Table 4: Percentage of healthy, MI cases or censored individuals in Model 0

MI indicator	%
Healthy	44
MI cases	27
Censored at age 80	29
Prevalent cases of all MI cases	6

4.2 Results from Model 1

The results from Model 1 is presented in Table 5. There are the calculated "real" value on β which is set to be the same as the means of all 1 000 $\hat{\beta}$ for Cox regression analysis. We also have the mean of all 1 000 $\hat{\beta}$ for case-cohort analysis. In the table we also can see the mean of the standard errors, se $\hat{\beta}$, from the two analyzing methods, then the calculated standard deviation, sd $\hat{\beta}$, and the 95% coverage for both models.

Table 5: Results from Model 1
In the Cox regression model there are 20 000 individuals, but in the case-cohort analysis the subcohort is only 2 000 individuals. The true value of β =mean$\hat{\beta}_{Cox}$.

Analysis	mean $\hat{\beta}$	mean se$\hat{\beta}$	sd$\hat{\beta}$	95% coverage
Cox regression model	0.2784	0.0232	0.0226	0.952
Case-cohort	0.2616	0.1074	0.1103	0.941
Case-cohort with prevalent	0.2700	0.1037	0.1066	0.942

Table 6: Mean square error, variance and bias for Model 1

Analysis type	mean square error	variance	bias
Case-cohort without prevalent cases	0.0125	0.0122	-0.0168
Case-cohort with prevalent cases	0.0114	0.0114	-0.0084

We could also be interested in knowing the percentage of individuals who are healthy until death, who have an MI before they die or who are censored because

they still live and are older than the censoring age, 80 years. These results we can see in Table 7.

Table 7: Percentage of healthy, MI cases or censored individuals in Model 1

MI indicator	%
Healthy	43
MI cases	28
Censored at age 80	29
Prevalent cases of all MI cases	12

5 Conclusions and discussion

The results do not indicate any strong selection bias by including the prevalent cases compared to excluding the prevalent cases in the case-cohort analysis.

The results for both models are more or less the same, see Table 3 and 6. The variance, or the precision, for the two models is almost the same if we include the prevalent cases or not. We might see a small difference that the variance is smaller when we include the prevalent cases than when we exclude them. The bias is larger when we exclude the prevalent cases than when we include them. But this difference is much smaller than the variance, so it is not visible in the mean square error.

We have, in Table 2 and 5, the results of the coefficient for the covariate β. When we do the analysis with the Cox regression model we receive the value that we call true, because all 20 000 individuals are being used in this study. The case-cohort design with and without the prevalent cases gives us β's lower than the true value. We can see that the mean of the standard error, mean se $\hat{\beta}$, and standard deviation, sd $\hat{\beta}$, is almost the same as they should be, if the standard error is correctly programmed in `coxph` and `cch` in the software R. The mean standard error for case-cohort design is so much larger than the mean standard error for the Cox regression model because in the Cox regression model we have more individuals, 20 000 individuals compared with 2 000 individuals in the case-cohort analysis. The values of the 95% coverage are for Model 0 slightly higher than 95 % while for Model 1 is almost 94%, see Table 2 and 5. This indicates that the estimated β's, $\hat{\beta}$, and their standard errors, se $\hat{\beta}$, are close enough to the real value in Model 0, but further away in Model 1.

If we would like to examine further if there is a difference between including or excluding the prevalent cases in the cases-cohort design we could create more prevalent cases. In this thesis 6 % and 11 % of all the MI cases are prevalent cases in Model 0 respective Model 1, see Table 4 and 7. We could also create more MI cases so that more than 27-28 % of all individuals have an MI, as can be seen in the same tables as mentioned above. This we could do to test the model with and without prevalent cases even if the data will get unrealistic by a higher rate of MI cases. Another way to is to create a larger difference between death after an MI. More data could also help us to see the difference between including or excluding the prevalent cases.

To make the generated populations more realistic, we could as covariates have the different genotypes, that we expect contributes to an MI. We could also have sex as a covariate.

A Data

In Table 8 you see what the data, generated by my R code, look like.

Table 8: Data*The 10 first data generated by Model 0*

No	genotype	age_death	MI	age_MI_final	subcohort
1	1	76.85544	0	NA	0
2	1	51.25060	0	NA	0
3	2	63.73509	1	52.47947	0
4	1	67.76001	1	58.56679	0
5	0	74.01712	0	NA	0
6	2	75.39422	0	NA	0
7	0	68.97590	0	NA	0
8	1	70.88697	0	NA	0
9	1	66.95569	0	NA	0
10	0	83.23083	2	NA	1

B R code

B.1 Genotypes

```
a1=rbinom(n,1,0.5)

a2=rbinom(n,1,0.5)

genotype = a1+a2 # gives us 0, 1, 2 represent aa, Aa, AA
```

B.2 Natural death - Healthy to death

```
p=runif(n,0,1)

epsilon=log(-log(1-p))

death_mean=78

sigma=1/9

alpha=log(death_mean/gamma(1+sigma))

death=exp(alpha + sigma*epsilon)
```

B.3 Model 0

```
# number of persons in my simulation
n=20000
############ From state Healthy to state MI in model 0 ###############
p=runif(n,0,1)
epsilon_1=log(-log(1-p))
mean_MI_aa=90
sigma=1/8
alpha_1=log(mean_MI_aa/gamma(sigma+1))
mean_MI_Aa=75
beta_1=log(mean_MI_Aa/mean_MI_aa)
age_MI=exp(alpha_1 + beta_1 * genotype + sigma*epsilon_1)
############ From state MI to state Dead in model 0 ###############
p=runif(n,0,1)
epsilon_2=log(-log(1-p))
mean_t_young=10
alpha_2=log(mean_t_young/gamma(sigma+1))
death_MI=age_MI+exp(alpha_2+sigma*epsilon_2)
###################### Create our data ########################
age_death=(death >= death_MI)*death_MI+(death < death_MI)*death
age_censoring=80
# 0=death, 1=MI, 2=censored by age
MI=(death<death_MI & death<age_censoring)*0 +
(death>=death_MI & death_MI<age_censoring)*1 +
(death>=age_censoring & death_MI>=age_censoring)*2
age_MI_final=((death<death_MI & death<=age_censoring)| #| for "or"
```

```
(death>age_censoring & death_MI>age_censoring))*(-1)+
(death>death_MI & death_MI<age_censoring)*age_MI
for(j in 1:n){
if(age_MI_final[j]==-1)
age_MI_final[j]=NA}
nr=1:n
subcohort=floor(nr/10)==nr/10
matrix_death=cbind(nr,genotype, age_death,MI,age_MI_final, subcohort)
data_death_model0=as.data.frame(matrix_death)
ratio_0=sum(data_death_model0$MI==0)/n
ratio_1=sum(data_death_model0$MI==1)/n
ratio_2=sum(data_death_model0$MI==2)/n
```

B.3.1 Model 0 analysis

```
library(survival)
# creating a data from Model 0 but only with the MI-cases
data_MI_model0=data_death_model0[MI==1,]
######## Cox analysis ############
Cox_model0=coxph(Surv(age_MI_final,MI==1)~genotype, data=data_death_model0)
######## Case-cohort analysis ############
## not prevalent cases
data_H_baseline=data_death_model0[(age_death>45)&(age_MI_final>45)&(MI==1),]
casecohort_result_model0=cch(Surv(data_H_baseline$age_MI_final) ~
data_H_baseline$genotype, data=data_H_baseline, subcoh=~subcohort,
id=~nr,cohort.size=n)
## include prevalent cases
data_H_baseline=data_death_model0[(age_death>45)&(MI==1),]
```

```
casecohort_result_model0_prevalent=cch(Surv(data_H_baseline$age_MI_final)
~data_H_baseline$genotype, data=data_H_baseline, subcoh=~subcohort,
id=~nr, cohort.size=n)
```

B.3.2 Model 0 looping

```
# number of loops
k=1000
for(ij in 1:k)
{source("Model 0 analysis.R")
# Cox
Cox_result_model0[ij]=coef(Cox_model0)
se_Cox_result_model0[ij]=sqrt(Cox_model0$var)
# case-cohort
casecohort_coef_model0[ij]=coef(casecohort_result_model0)
casecohort_se_model0[ij]=sqrt(casecohort_result_model0$var)
# case-cohort prevalent cases included
casecohort_coef_model0_prevalent[ij]=coef(casecohort_result_model0_prevalent)
casecohort_se_model0_prevalent[ij]=sqrt(casecohort_result_model0_prevalent$var)
# the rates
ratio_0_all_model0[ij]=ratio_0
ratio_1_all_model0[ij]=ratio_1
ratio_2_all_model0[ij]=ratio_2
} # end of for-loop

### Cox results ###
mean_Cox_result_model0=mean(Cox_result_model0,na.rm=TRUE)
mean_se_model0=mean(se_Cox_result_model0)
```

```
sd_Cox_model0=sd(Cox_result_model0)
### Case-cohort results ###
mean_casecohort_coef_model0=mean(casecohort_coef_model0)
mean_se_cch_model0=mean(casecohort_se_model0)
sd_cch_model0=sd(casecohort_coef_model0)
### Case-cohort results prevalent cases included ###
mean_casecohort_coef_model0_prevalent=mean(casecohort_coef_model0_prevalent)
mean_se_cch_model0_prevalent=mean(casecohort_se_model0_prevalent)
sd_cch_model0_prevalent=sd(casecohort_coef_model0_prevalent)
###### 95% coverage ######
# true beta
beta_input=mean_Cox_result_model0
for (ik in 1:k)
{procent95[ik]=1.96*se_Cox_result_model0[ik]
analys95proc_Cox[ik]=((beta_input>=(Cox_result_model0[ik]-procent95[ik]))&
(beta_input<=(Cox_result_model0[ik]+procent95[ik])))
procent95_cch[ik]=1.96*casecohort_se_model0[ik]
analys95proc_cch[ik]=((beta_input>=(casecohort_coef_model0[ik]-procent95_cch[ik]))&
(beta_input<=(casecohort_coef_model0[ik]+procent95_cch[ik])))
procent95_cch_prevalent[ik]=1.96*casecohort_se_model0_prevalent[ik]
analys95proc_cch_prevalent[ik]=(
(beta_input>=(casecohort_coef_model0_prevalent[ik]-procent95_cch_prevalent[ik]))&
(beta_input<=(casecohort_coef_model0_prevalent[ik]+procent95_cch_prevalent[ik])))
} # end for-loop
percent_beta_input_Cox=mean(analys95proc_Cox)
percent_beta_input_cch=mean(analys95proc_cch)
percent_beta_input_cch_prevalent=mean(analys95proc_cch_prevalent)
```

```
# rates
mean(ratio_0_all_model0)
mean(ratio_1_all_model0)
mean(ratio_2_all_model0)
# MSE
Bias_model0_without_prevalent=mean_casecohort_coef_model0-mean_Cox_result_model0
Bias_model0_with_prevalent=mean_casecohort_coef_model0_prevalent-mean_Cox_result_model0
var_model0_without_prevalent=var(casecohort_coef_model0)
var_model0_with_prevalent=var(casecohort_coef_model0_prevalent)
MSE_model0_without_prevalent=var_model0_without_prevalent+
Bias_model0_without_prevalent^2
MSE_model0_with_prevalent=var_model0_with_prevalent+
Bias_model0_with_prevalent^2
rate_prevalent_cases=sum((age_MI<45) & (MI==1))/
sum(data_death_model0$MI==1);rate_prevalent_cases
```

B.4 Model 1

```
# number of persons in my simulation
n=20000
############## Simulate age for MI - model 1 #################
p=runif(n,0,1)
epsilon_prel=log(-log(1-p))
mean_MI_aa=90
sigma=1/8
alpha_1=log(mean_MI_aa/gamma(sigma+1))
mean_MI_Aa=75
beta_1=log(mean_MI_Aa/gamma(1+sigma))-alpha_1
```

```
age_MI_prel=exp(alpha_1 + beta_1 * genotype + sigma*epsilon_prel)
## For this model we need to draw the age_MI_prel for all age_MI_prel>=45
p=runif(n,0,1)
epsilon_1=log(-log(1-p))
age_MI=age_MI_prel*(age_MI_prel<45)+
exp(alpha_1 + beta_1 * genotype + sigma*epsilon_1)*(age_MI_prel>=45)
############ From state MI to state Dead ################
t_before45=5 #years
t_after45=10
beta_2=log(t_after45/t_before45)
alpha_2=log(t_before45/gamma(sigma+1))
p=runif(n,0,1)
epsilon_2=log(-log(1-p))
TorF=age_MI>=45
death_MI=age_MI+
exp(alpha_2+beta_2*TorF+sigma*epsilon_2)
###################### Create our data ########################
age_death=(death >= death_MI)*death_MI+(death < death_MI)*death
age_censoring=80
# 0=death, 1=MI, 2=censored by age
MI=(death<death_MI & death<age_censoring)*0 +
(death>=death_MI & death_MI<age_censoring)*1 +
(death>=age_censoring & death_MI>=age_censoring)*2
age_MI_final=((death<death_MI & death<=age_censoring)| #| for "or"
(death>=age_censoring & death_MI>=age_censoring))*(-1)+
(death>=death_MI & death_MI<age_censoring)*age_MI
for(j in 1:n){ if(age_MI_final[j]==-1) age_MI_final[j]=NA}
```

```
nr=1:n
subcohort=floor(nr/10)==nr/10
matrix_death=cbind(nr,genotype, age_death,MI,age_MI_final,subcohort)
data_death_model1=as.data.frame(matrix_death)
```

B.4.1 Model 0 analysis

```
######## Cox analysis ############
Cox_model1=coxph(Surv(age_MI_final,MI==1)~genotype, data=data_death_model1)
######## Case-cohort analysis ############
data_H_baseline=data_death_model1[(age_death>45)&(age_MI_final>45)&(MI==1),]
# excluding the prevalent cases
casecohort_result_model1=cch(Surv(age_MI_final) ~ genotype,
data=data_H_baseline, subcoh=~subcohort, id=~nr, cohort.size=n)
# data WITH the prevalent cases
data_H_prevalent=data_death_model1[(age_death>45)&(MI==1),]
casecohort_result_model1_prevalent=cch(Surv(age_MI_final) ~ genotype,

data=data_H_prevalent, subcoh=~subcohort, id=~nr, cohort.size=n)
```

B.4.2 Model 1 looping

```
# number of loops
k_1=1000
for(i_1 in 1:k_1) # start of for-loop
{source("Model 1 analysis.R")
Cox_result_model1[i_1]=coef(Cox_model1)
se_Cox_result_model1[i_1]=sqrt(Cox_model1$var)
casecohort_coef_model1[i_1]=coef(casecohort_result_model1)
```

```
casecohort_se_model1[i_1]=sqrt(casecohort_result_model1$var)
casecohort_coef_model1_prevalent[i_1]=coef(casecohort_result_model1_prevalent)
casecohort_se_model1_prevalent[i_1]=sqrt(casecohort_result_model1_prevalent$var)
## the rates
ratio_0_all_model1[i_1]=ratio_0_model1
ratio_1_all_model1[i_1]=ratio_1_model1
ratio_2_all_model1[i_1]=ratio_2_model1} # end of for-loop

mean_Cox_result_model1=mean(Cox_result_model1,na.rm=TRUE)
mean_se_model1=mean(se_Cox_result_model1)
sd_Cox_model1=sd(Cox_result_model1)
mean_casecohort_coef_model1=mean(casecohort_coef_model1)
mean_casecohort_se_model1=mean(casecohort_se_model1)
sd_casecohort_coef_model1=sd(casecohort_coef_model1)
mean_casecohort_coef_model1_prevalent=mean(casecohort_coef_model1_prevalent)
mean_casecohort_se_model1_prevalent=mean(casecohort_se_model1_prevalent)
sd_casecohort_coef_model1_prevalent=sd(casecohort_coef_model1_prevalent)
###### 95% coverage ######
beta_input_model1=mean_Cox_result_model1
for (ik in 1:k_1)
{procent95_model1[ik]=1.96*se_Cox_result_model1[ik]
analys95proc_Cox_model1[ik]=((beta_input_model1
>=(Cox_result_model1[ik]-procent95_model1[ik]))&
(beta_input_model1<=(Cox_result_model1[ik]+procent95_model1[ik])))
procent95_cch_model1[ik]=1.96*casecohort_se_model1[ik]
analys95proc_cch_model1[ik]=(
(beta_input_model1>=(casecohort_coef_model1[ik]-procent95_cch_model1[ik]))&
```

```
(beta_input_model1<=(casecohort_coef_model1[ik]+procent95_cch_model1[ik])))
procent95_cch_model1_prevalent[ik]=1.96*casecohort_se_model1_prevalent[ik]
analys95proc_cch_model1_prevalent[ik]=(
(beta_input_model1>=
(casecohort_coef_model1_prevalent[ik]-procent95_cch_model1_prevalent[ik]))&
(beta_input_model1<=
(casecohort_coef_model1_prevalent[ik]+procent95_cch_model1_prevalent[ik])))
} # end for-loop
percent_beta_input_Cox_model1=mean(analys95proc_Cox_model1)
percent_beta_input_cch_model1=mean(analys95proc_cch_model1)
percent_beta_input_cch_model1_prevalent=mean(analys95proc_cch_model1_prevalent)
```

C References/Bibliography

References

[1] *http://www.r-project.org/.*

[2] Myocardial infarction. www.wikipedia.org, 21 Sept 2007.

[3] Single nucleotide polymorphism. www.wikipedia.org, 5 March 2008.

[4] Weibullfördelning. www.wikipedia.se, 1 June 2008.

[5] Dorothy D. Dunlop Ajit C. Tamhane. *Statistics and data analysis from elementary to intermediate*, page 23. Prentice Hall, 2000.

[6] Dorothy D. Dunlop Ajit C. Tamhane. *Statistics and data analysis from elementary to intermediate*, page 49. Prentice Hall, 2000.

[7] Dorothy D. Dunlop Ajit C. Tamhane. *Statistics and data analysis from elementary to intermediate*, pages 198–199. Prentice Hall, 2000.

[8] Jonny Österman Carl Nordling. *Physics Handbook for Science and Engineering*, page 417. Studentlitteratur, 2006.

[9] D. Collett. *Modelling survival data in medical research.* Chapman & Hall/CRC, 1994.

[10] Kulathinal S Asplund K Cambien F Ferrario M Perola M Peltonen L Shields D Tunstall-Pedoe H Kuulasmaa K Evans A, Salomaa V. Morgam (an international pooling of cardiovascular cohorts). *Int J Epidemiol*, 34:21–27, 2005.

[11] Joanna Tyrcha/Karin Fremling. Lecture notes from the course stochastic processes and simulation 1.

[12] Håkon K. Gjessing Odd.O Aalen, Ørnulf Borgan. *Event History Analysis A Process Point of View.* Springer, Preliminary version September 2007.

[13] Theadore Colton Peter Armitage, editor. *Encyclopedia of biostatistics, reprinted 1999*, pages 497–503. John Wiley & Sons Ltd, 1998.

[14] Larry Goldstein Janice Pogoda Ørnulf Borgan, Sven Ove Samuelsen. Exposure stratified case-cohort design. *Lifetime Data Analysis*, 6:39–58, 2000.

[15] M.R. Cummings W.S Klug. *Essentials of genetics, 5th edition*, chapter Glossary. Pearson Education, Inc., 2005.

[16] M.R. Cummings W.S Klug. *Essentials of genetics, 5th edition*, page 214. Pearson Education, Inc., 2005.