THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

PROGNOSTIC FUNCTIONS BASED ON MULTI-STATE MODELS

By

DIMITRE STEFANOV

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2007

The members of the Committee approve the Dissertation of Dimitre Stefanov defended on Aug 10, 2007.

Dan McGee (Sr)
Professor Directing Dissertation

Isaac Eberstein
Outside Committee Member

Fred Huffer
Committee Member

Xufeng Niu
Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.

To my parents Georgi Dimitrov Stefanov and Rositsa K. Stefanova

ACKNOWLEDGEMENTS

I am deeply indebted to my major professor, Dr. Dan McGee for his guidance, patience and encouragement. His help was essential for the completion of my dissertation. I have learned a lot from him and I wish I listened better and learned even more from his advice. My gratitude also goes to Dr. Huffer and Dr. Niu for all the help and discussions during the years.

I came to the Department with Math and Theoretical Statistics background. I naturally had a tendency to focus on the technique, rather than on the problem. Most of the classes I took after I joined the Statistics Department were from the members of my dissertation committee. In these classes I have learned from Dr.McGee and Dr.Niu how to focus on a particular problem with its specific details. I have learned from Dr. Huffer to look for an interpretation which leads to developing an intuition and deeper understanding of the theory. I want to thank Dr. Carlson and Dr. Eberstein for serving in my committee and for their comments showing me different angles to look at my research.

Radha has helped me enormously for my teaching and I learned a lot from her how to be an effective teacher. Dr. Ramsier has made very important contributions to my teaching skills giving me very constructive feedback and sharing his experience and ideas with me.

My appreciation extends to Jennifer and Pam for always being helpful, knowledgeable and good spirited. James has always been available and ready to help and explain what the problem is, no matter whether it is very simple or very complicated one.

Over the years I spent working on my dissertation there have been many people who made me feel at home in Tallahassee and in the Department. A list of all of them would be impossible, and I am afraid I will inevitably omit somebody. However, three of them has really made a difference for my experience in Tallahassee. I want to thank Mahtab, Dennis and Dimitre Tsigantchev for being great friends.

I have spent many wonderful hours with Radu, Wendy, Radha, Fei, Dai Ho and Shuva. The Bulgarian community in Tallahassee is really remarkable and I was fortunate to have good friends, with whom we shared a lot of good times and good food (e.g. Bulgarian cheese).

Finally I wish to express my deepest gratitude to my parents for their love and understanding during these years. My father particularly has been the greatest support, unconditionally helping me in everything I tried in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Multi-state models are models for a process, which at any time occupies one of several possible states. An example of a multi-state process is the life history of an individual, where the states can be different diseases and an absorbing state-death. We applied these methods to study cardiovascular diseases (CVD) and how they affect mortality. With the increasing proportion of elderly people in most developed countries, the burden of CVD on the society is increasing as well. It is estimated that by year 2020 heart disease and stroke will become leading cause of death and disability world wide. The number of fatalities is projected to increase to more than 20 million a year, and more than 24 million by year 2030. (Atlas of Heart Disease and Stroke, WHO, September 2004)

Prognostic models have been widely used by clinicians to predict the outcomes for patients free of CVD. These models have been developed mainly using risk functions for the binary outcome (yes=CVD, no=no CVD) in logistic regression or for modelling the failure time (time to death) in survival analysis. In both approaches, the focus is to determine the effect of the covariates (fixed at baseline or time-varying) to mortality. As the population ages and more people experience different diseases or events, such as heart attack or stroke, which do irreversible damage to the heart/brain and change the life expectancy. It is also expected, that factors like high blood pressure or diabetes may have different effects for a person before and after a stroke. The question that we are interested is how to model the event history for individuals who go through different disease states in their lifetime. The goal is to include information for a set of covariates as well as the time and the type of disease people encounter. We approach this problem from a multi-state prospective, where the states describe the progression of the disease, for example healthy state, coronary heart disease (CHD state) cerebral vascular accident (stroke) and death (absorbing state). The

problem can be generally divide into steps:

The first step is to estimate how transition rates between various states depend on the covariates. This will allow us to compare the role of covariates for different transitions.

The second step is to combine the estimated rates for a given set of covariates into appropriate transition rates. This will allow us to calculate a survival probability for a given subject. This can be used as a prognostic function at baseline, as well as at a later time, when information for the event history of the subject is available.

CHAPTER 1

Introduction

1.1 Brief history of prognostic models for CVD

Prognostic models for the Cardiovascular Disease (CVD) have been used widely. The focus has been on modeling time to death or time to the development of the disease. The standard approach has been to include possibly other diseases as covariates, recorded at baseline along with some prognostic factors (such as blood pressure, cholesterol). However, with the advances in medicine and health care in the second half of the 20th century, life expectancy is continually increasing with a large proportion of the population in their 80's and 90's. Many of the elderly people experience some form of Cardiovascular Disease and often live with it for a substantial amount of time. These diseases often leave people with irreplaceable damage to the heart (myocardial infarction) or to the brain (CVA). This phenomena suggests modeling the life history of the individuals including more states, such as CHD or CVA. First I will review the methods that have been used for building prognostic models. I will briefly outline the logistic regression, the Cox proportional hazards model, and some parametric models, all used in the context of the two state model, i.e. we consider state 0 - when people are alive and we follow them till failure (state 1) or censoring.

1.1.1 Logistic Regression

In logistic regression, we are interested in modeling the probability that a subject with given set of covariates will fail within a specified amount of time. We are not interested of the exact time of failure, only in the outcome at the end of the follow up period.

Suppose we have followed n individuals for a period of time t and recorded their covariates at the beginning of the study and whether they have failed by the end of the study. The response variable Y is binary - takes the value 1 if the subject has failed , and 0 if he/she

hasn't failed.

Let us denote

$\boldsymbol{X} = (1, x_1, \ldots, x_p)$ - the covariate vector and

$\pi(\boldsymbol{X}) = P(Y = 1|\boldsymbol{X}) = 1 - P(Y = 0|\boldsymbol{X})$ - the probability of failure by the end of the follow up.

The logistic regression model has the form:

$$logit(\pi(\boldsymbol{X})) = log\left(\frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})}\right) = \boldsymbol{X}\boldsymbol{\beta} \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ are the regression coefficients that require estimation.

The interpretation for the coefficients is given in terms of the odds ratio. Let us assume diabetes is one of the variables recorded at the beginning of the study, with the value 1 for those with diabetes and 0 otherwise. Then the odds ratio for two imaginary subjects, one with diabetes and the other without (having all the other covariates the same) is $\exp(\beta_{diab})$. The interpretation of the regression coefficient β_j for a continuous variable X_j is similar, comparing the odds ratio for a unit increase in X_j.

The regression coefficients are estimated to maximize the likelihood. Since the observations are assumed to be independent, the likelihood function is the product:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{X}_i)^{Y_i}(1 - \pi(\boldsymbol{X}_i))^{1-Y_i} \tag{1.2}$$

It is easier to maximize the log likelihood function

$$L(\boldsymbol{\beta}) = log(l(\boldsymbol{\beta})) = \sum_{i=1}^{n} Y_i log(\pi(\boldsymbol{X}_i)) + (1 - Y_i) log(1 - \pi(\boldsymbol{X}_i)) \tag{1.3}$$

To find the vector $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$, the first derivatives with respect to β_i $i = 1, \ldots, p$ are set to zero, obtaining the so called likelihood or score equations:

$$\sum_{i=1}^{n} \boldsymbol{X_i}(Y_i - \pi(\boldsymbol{X_i})) = 0 \tag{1.4}$$

which is a vector equation with $(p + 1)$ elements. Using the fact that the first coordinate in every vector $\boldsymbol{X_i}$ is one, we obtain the equality

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \pi(\boldsymbol{X_i}) \tag{1.5}$$

which indicates that the number of people expected to fail is the same as the number of people observed to fail. There is no closed form for the coefficients maximizing the likelihood and numerical methods such as Newton-Raphson can be used.
Logistic regression has been widely used in prognostic models [1], [2]. However, this approach has the limitations that all subjects are observed over the same period of time and it estimates the survival function only at one specified time.

1.1.2 Cox Proportional Hazards Model

The Cox model [3] assumes the hazard rate as follows:

$$\alpha(t|\boldsymbol{X}) = \alpha_0(t)exp(\boldsymbol{X}\boldsymbol{\beta}) \tag{1.6}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to be estimated and $\boldsymbol{X}$ is a vector with covariates. If we compare the hazards ratio for subjects i and j with covariates vectors $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$, we have:

$$\frac{h(t|\boldsymbol{X_i})}{h(t|\boldsymbol{X_j})} = exp(\boldsymbol{\beta}'(\boldsymbol{X_i} - \boldsymbol{X_j}))$$

which does not depend on time. Because of this, the Cox model is often called the proportional hazards model.
If the covariate vectors $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ differ in only one variable, say subject i is 55 year old and subject j is 54 then the hazard ratio is $\exp(\beta_{age})$. The exponentiated individual coefficients can be interpreted as the hazards ratio for a unit increase in the corresponding covariate.
The baseline hazard $\alpha_0(t)$ only depends on time t and can vary freely -hence the Cox model is semi-parametric. The strength of the Cox model is the ability to estimate the coefficients $\boldsymbol{\beta}$ without specifying the baseline hazard $\alpha_0(t)$. This is achieved by maximizing the partial likelihood, which has contributions for each death time.
Let $t_1, t_2, \ldots t_D$ denote the ordered event times. Let us assume there are no tied death times and the censoring is independent of the failure times. Let $\boldsymbol{X}_{(j)}$ be the covariate vector for the subject failing at time t_j. The partial likelihood is defined:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{X}_{(i)}\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{X_j}\boldsymbol{\beta})}$$

where $R(t_j)$ is the risk set at time t_j -the set of all individuals who are still under observation at a time just prior to t_j. The partial likelihood depends only on the order of the failure

times, not on the exact times.

Wilson et. al. [4] used the Cox proportional hazards model for the probability of developing CHD using data from the Framingham Study. The covariates included in the model are age, systolic blood pressure, cigarette use, diabetes, total cholesterol (or LDL cholesterol) and HDL cholesterol. The continuous covariates are split into intervals and a model with only categorical variable is considered. The discriminatory ability of the model is compared to proportional hazards model and accelerated failure model with continuous covariates using the ROC curve and the corresponding c statistics. The performance of these models is very similar.

1.1.3 Parametric models

In general, parametric models are written in two different ways:

- proportional hazards models- specifying the hazard rate

- accelerated failure-time (AFT) models - specifying directly the failure time T.

1.1.4 Accelerated Failure Time models

In this approach, the failure time T is modeled directly:

$$ln(T_j) = \boldsymbol{X_j\beta} + ln(\tau_j)$$

where $\tau_j = \exp(-\boldsymbol{X_j\beta})T_j$ is a random variable with a distribution assumed to be known. The quantity $\exp(-\boldsymbol{X_j\beta})$ is called the acceleration parameter. The role for the acceleration parameter can be viewed through the survival function.

Let us denote $S_0(t)$ the survival function of a subject with a covariate vector $\boldsymbol{X} = \boldsymbol{0}$. Then we have

$$S(t|\boldsymbol{X}) = S_0(\exp(-\boldsymbol{X_j\beta})t)$$

In other words, the probability of survival after time t for a subject with covariates $\boldsymbol{X}$ is equivalent to the probability of survival past time $\exp(-\boldsymbol{X_j\beta})t$ for a subject with all covariate values equal to zero. Therefore, the interpretation of the acceleration parameter is the following:

- if $\exp(-\boldsymbol{X_j\beta}) = 1$, then time passes at its normal rate
- if $\exp(-\boldsymbol{X_j\beta}) > 1$, then time is accelerated and failure is expected to occur sooner
- if $\exp(-\boldsymbol{X_j\beta}) < 1$, then time passes more slowly and failure is expected to occur later.

Generally, the two approaches described above are different ways to model the dependence on the covariates for the survival function. However, if the Weibull distribution is assumed for the random variable $\tau_j = \exp(-\boldsymbol{X_j\beta})T_j$, then it can be shown [5] that the regression variables act multiplicatively on the hazard function. Therefore, using the Weibull distribution, both approaches lead to the same model and this is the only distribution with this property.

Anderson [6], [7] generalizes the AFT models using the Weibull distribution. To illustrate his approach we need to point out two observations. First is an alternative definition for the 2 parameter Weibull distribution:
A random variable T has Weibull distribution with a location parameter μ and scale parameter σ if.

$$U = \frac{log(T) - \mu}{\sigma} \tag{1.7}$$

has an extreme value distribution with cdf $F_U(u) = 1 - \exp(-\exp(u))$.
The standard definition of the Weibull distribution [5] has hazard function :
$\alpha(t) = \lambda\gamma(\lambda t)^{(\gamma-1)}$
The two definitions specify the same distribution with the correspondence:
$\gamma = \frac{1}{\sigma}$ and $\lambda = \exp(-\mu)$.
The advantage of the definition through the parameters μ and σ is that the standard AFT model, with Weibull distribution as a baseline hazard, is equivalent to modeling the location parameter μ as a function of the covariates:
$\mu(\boldsymbol{X}) = \boldsymbol{X\beta}$ and σ considered a constant.

The second observation, is that the property of the proportional hazards model 1.1.2, can be expressed using the logarithm of the cumulative hazard function, defined at the beginning of Chapter 2:

$$log(A(t|\boldsymbol{X_i})) - log(A(t|\boldsymbol{X_j})) = \boldsymbol{\beta}'(\boldsymbol{X_i} - \boldsymbol{X_j})$$

in other words, the difference of the log cumulative hazards is constant with time.

In the AFT Weibull model, $\mu = \mu(\boldsymbol{X})$ and the logarithm of the cumulative baseline hazard is $A(t|\mu, \sigma)$:

$$log(A(t|\mu, \sigma)) \equiv log(-log(Pr\{T \geq t\})) = \frac{log(t) - \mu}{\sigma}$$

where the first equality is a general formula (which will be outlined in Chapter 2) and the second equality uses the definition of the Weibull distribution 1.7. If σ is treated constant, as in the general AFT Weibull model, the difference of the log cumulative hazards for subjects i and j with covariates $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ is a constant:

$$log(A(t|\boldsymbol{X_i})) - log(A(t|\boldsymbol{X_j})) = \frac{\mu(\boldsymbol{X_i}) - \mu(\boldsymbol{X_j})}{\sigma}$$

and it is a proportional hazards model.
Anderson suggests using models, where σ is also a function of the covariates. He compares the following models, assuming underlying Weibull distributions as in (1.7) :

- Model 1: $\mu = \boldsymbol{X\beta}$, $log(\sigma) = \theta_0$
- Model 2: $\mu = \boldsymbol{X\beta}$, $log(\sigma) = \theta_0 + \theta_1\mu$
- Model 3a: $\mu = \boldsymbol{X\beta}$, $log(\sigma) = \theta_0 + \theta_1\mu + \theta_2\mu^2$
- Model 3b: $\mu = \boldsymbol{X\beta}$, $log(\sigma) = \boldsymbol{X\gamma}$

Model 1 is the standard AFT Weibull model. Model 2 is the final model, where different models are compared using the Likelihood Ratio Test. The parameter θ_1 indicates how fast the logarithm of the cumulative hazard converges or diverges for two values of μ. The covariates used in the models are measured at baseline: logarithm of age, logarithm of systolic blood pressure, logarithm of cholesterol, Metropolitan relative weight (Metropolitan Life Insurance Company Tables) and an indicator for smoking. The use of logarithms resulted in improved fit. The final model was used to predict the probability of developing CHD for a given set of covariates.

1.1.5 Parametric Proportional Hazards models

In the models in this category, the hazard is specified as in the Cox model,

$$\lambda(t|\boldsymbol{X_j}) = \lambda_0(t)\exp(\boldsymbol{X_j\beta})$$

where $\boldsymbol{X_j}$ are the covariate vector for subject j, and $\boldsymbol{\beta}$ is the vector for the regression coefficients. The difference from the Cox model is that parametric assumptions are made about the baseline hazard function. Conroy et al. [8] used risk functions based on a Weibull proportional hazards model, stratified on cohort and gender. The SCORE (Systematic Coronary Risk Evaluation) project uses data from 12 European cohort studies and the goal is to measure the risk of fatal cardiovascular disease. The approach of measuring the risk of fatal CVD disease is in contrast to most other studies, where the end point is fatal or non-fatal CVD disease. The estimation of the risk function is split into two parts-the risk of fatal CHD disease and the risk of fatal non -CHD CVD disease. One reason for this distinction is the well observed regional differences in Europe- countries with high CVD risk, such as Denmark, Finland and Norway and low CVD risk regions - Belgium, Italy and Spain. In the countries with low total CVD mortality, the proportion of the CHD deaths to the total CVD deaths is also low. The model uses strata for cohorts and gender, allowing the baseline hazards to vary, but keeping the coefficients for the risk factors the same. In other words, the model assumes that the effect of the covariates is the same for all individuals. Another notable difference from other studies is the use of age as a time scale. This idea was first used by Korn et al. [9], where they have a discussion for the use of age as a time scale for certain (e.g. chronic) diseases. The covariates considered in the SCORE study are smoking status, systolic blood pressure and cholesterol (or cholesterol to HDL ratio). Information for diabetes was not available for all the cohorts and was omitted. Two models were considered, one using total cholesterol, the other the ratio of total cholesterol and HDL with very similar results. The risk estimations are displayed graphically in risk charts for the use by clinicians.

Non-parametric and semiparametric methods compare subjects at the time subjects fail. Parametric models on the other hand, calculate probabilities for every interval, for every subject. The likelihood function, in both of the parametric models is the product of the individual contributions for each subject, for the time they were under observation. If subject j was followed in the interval $(t_{0j}\ t_j]$, his contribution to the likelihood is :

$$L_j(\Theta) = \frac{S(t_j|\boldsymbol{X_j}, \Theta)^{1-d_j} f(t_j|\boldsymbol{X_j}, \Theta)^{d_j}}{S(t_{0j}|\boldsymbol{X_j}, \Theta)}$$

where d_j is an indicator for censoring ($d_j = 0$ if censored, $d_j = 1$ if failure), Θ is the vector of parameters used in the model and $f(t|\boldsymbol{X}, \Theta)$ is the density function of the failure time T.

Multi-state models describe the life history of individuals and are helpful in studying the

occurrence of diseases changing the risk of death. They can be used for modeling recurring events, although in this dissertation we focus on the first occurrence of diseases. Most of the studies for CVD developed risk functions based on end points of fatal and non-fatal CVD. There are a number of problems using non-fatal end points, one of which is that CVD is a combination of different diagnosis with different severity, as indicated in [8]. However, using CVD with fatal end points only, will inevitably include the burden of the non-fatal CVD events occurring earlier. Peeters [10] estimated using Multi-state life tables that at age 50 men have 6.3 (5.7 for women) years spent with CVD. We believe, a more complete picture would be to model the disease as a factor to mortality, adjusting for covariates. Modeling CVD diseases as states, we believe models a natural process of aging. De Becker et al [11], listed 5 observations on which the prevention of CVD is based in the SCORE project. The second observation indicates that "the underlying pathology is usually artherosclerosis, which develops insidiously over many years". Multi-state models can also be used for making estimates of individual prognosis in epidemiological or clinical studies, combining the estimated prognosis for developing a disease and the mortality afterwords.

CHAPTER 2

Background

2.1 Background on Survival Analysis

Survival analysis is an area which focuses on the distribution of a failure time T. The probability distribution of T can be specified in different ways: with the probability density function (pdf), the survival function or the hazard function. Each one determines the rest.

The survival function is defined as the probability that T exceeds a value t in its range: $S(t) = P(T > t)$. The distribution of the the time to failure T can be discrete, continuous and mixed, and we will briefly outline each of these cases.

If T is a discrete random variable taking values $t_1 < t_2, \ldots$ with associated probability function

$$f(t_i) = P(T = t_i),\ i = 1, 2, \ldots,$$

the hazard at t_i is defined as the conditional probability of failure at t_i given that the individual has survived to t_i:

$$\alpha_i = P(T = t_i | T \geq t_i) = \frac{f(t_i)}{S(t_i)},\ i = 1, 2, \ldots,$$

and the cumulative hazard is defined as

$$A(t) = \sum_{t_j \leq t} \alpha_j$$

The relation between the survival function and the hazard function is given with the formula

$$S(t) = \prod_{t_j \leq t} (1 - \alpha_j)$$

If T is a continuous random variable, the probability density function of t is defined as $f(t) = -dS(t)/dt$ and the intensity (hazard) function is defined as

$$\alpha(t) = \lim_{\Delta t \to 0} (\Delta t)^{-1} P(t \leq T < t + \Delta t | T \geq t) \tag{2.1}$$

It is the instantaneous rate at which failures occur for individuals that are surviving at time t. We have the following relation

$$S(t) = exp\left[-\int_0^t \alpha(s)ds\right] = exp\left[-A(t)\right] \tag{2.2}$$

where $A(t) = \int_0^t \alpha(s)ds$ is called the cumulative hazard function.
We will derive another formula for $P[T \in [t, t+dt) \,|\, T \geq t]$ and compare with the one above:

$$P[T \in [t, t+dt) \,|\, T \geq t] = \{S(t-) - S(t+dt)\}/S(t-)$$

which in the case of continuous failure time T, using (2.2), can be written as

$$1 - exp\{-(A(t+dt) - A(t))\} \approx dA(t)$$

The last approximation is based on Taylor's formula expansion $1 - e^{-x} \approx x$, for small values of x. More generally, the distribution of T may have both discrete and continuous parts. The cumulative hazard function is then:

$$A(t) = \sum_{t_j \leq t} \alpha_j + \int_0^t \alpha_c(s)ds$$

We can specify the hazard of failure over the infinitesimal interval $[t, t+dt]$ as

$$\begin{aligned} dA(t) &= A(t+dt) - A(t-) \\ &= P[T \in [t, t+dt) \,|\, T \geq t] \\ &= \begin{cases} \alpha_i & t = t_i,\ i = 1, 2, \ldots \\ \alpha_c(t)dt & otherwise \end{cases} \end{aligned}$$

The survival function is given with the formula:

$$S(t) = \exp\left\{-\int_0^t \alpha_c(u)du\right\} \prod_{t_j \leq t} (1 - \alpha_j)$$

There is a general formula for the survivor function that holds for discrete, continuous, or mixed cases

$$S(t) = \Pi_0^t (1 - dA(s))$$

where the product integral on the right hand side will be defined later in section 2.4.2.

2.2 Specifying the failure time distribution

2.2.1 Kaplan-Meier formula

In the case of a homogeneous population (when no covariates are considered), we have different ways to specify the distribution. A non-parametric estimate for the survival function is given by the Kaplan-Meier formula:

$$S(t) = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j}$$

where $t_1 < t_2 < \ldots < t_k$ represent the observed failure times, n_j is the number of people at risk and d_j is the number of deaths at time t_j. Later we will outline some parametric and non-parametric methods for a homogeneous population in the context of multi-state models. We have already discussed methods for specifying the distribution of a falure time adjusting for covariates. The Cox proportional hazards model and the accelerated failure time model were discussed in Chapter 1 and present a semiparametric and a parametric approach for modeling survival data. Here, we will add another (non-parametric) method.

2.2.2 Aalen Additive hazard model

This model originates from the work of Aalen [12] and is an alternative to the semiparametric multiplicative hazard model. We have, in the case of the two-state mortality model,

$$\alpha(t, \boldsymbol{X}) = \beta_0(t) + \beta_1(t)X_1(t) + \ldots + \beta_k(t)X_k(t),$$

where $\boldsymbol{X}(t) = [X_1(t), \ldots, X_k(t)]$ is a vector of possibly time dependent covariates. Use of this model for multi-state problems is presented in Aalen [13] and will be discussed later. The two main differences from the multiplicative hazards model are:
- the hazard of a failure time is modeled as a linear combination of the covariates
- a non-parametric approach is adopted, i.e. the coefficients $\beta_i(t)$, $i = 1, \ldots, k$ are allowed to vary freely (they are assumed constants in the proportional hazards Cox model).

As outlined by McKeague and Sasieni [14] the additive form can be interpreted loosely in terms of unobserved competing risks, since the hazard function for the minimum of independent random variables is the sum of the hazards for the individual variables.

2.3 Stochastic models

Multi-state processes decribe the life history of individuals. At any time the process occupies one state. We can view the classical survival analysis as a process with two states, alive (state 0) and dead (state 1). We will refer to this model as the mortality model. Next, we will illustrate the competing risk model.

The competing risk model

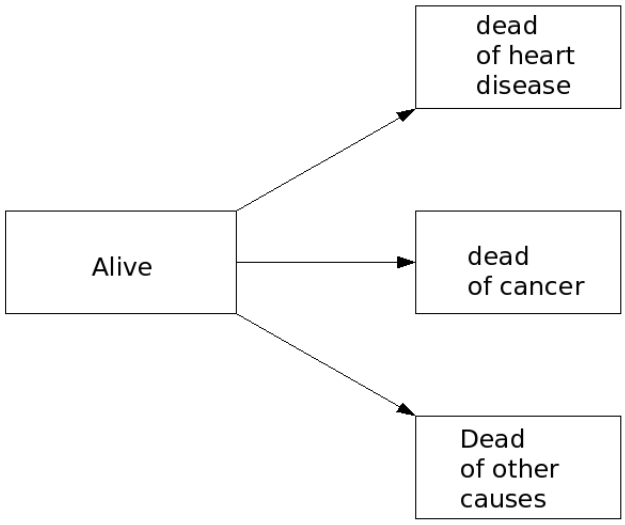

Figure 2.1: Competing Risk Model

Figure 2.1 presents the competing risk model. In this model, subjects are alive in state 0 and the other states are different causes of death. These models have been widely used and there is an extensive literature on them (for a review: Crowder [15]).
Denote k the number of competing causes of death. We have k possible transitions, each with a specific hazard rate $\alpha_{0j}(t)$, $j = 1, \ldots, k$. Denote $P_{0j}(s,t)$ the probability that the process is in state 0 at time s and in state j at time t, $j = 0, 1, \ldots, k$. The hazard rates are

defined as:

$$\alpha_{0j}(t) = \lim_{\Delta t \to 0} (\Delta t)^{-1} P_{0j}(t,\, t + \Delta t)$$

There are two types of transition probabilities, the survival function:

$$P_{00}(0,\, t) = S(t) = P(T > t) = exp\left(- \int_0^t \sum_{h=1}^{k} \alpha_h(u) d(u) \right)$$

and the cause specific cumulative incidence functions:

$$P_{0h}(0,\, t) = \int_0^t S(u) \alpha_h(u) d(u), \;\; h = 1, \ldots, k$$

Next, we will introduce the notation for a general multi-state model.

2.3.1 General notations and definitions

A multi-state process is a stochastic process $(X(t),\, t \in \Omega)$ with a finite state space $S = 1, \ldots, p$ and with right continuous piecewise constant sample paths, with limits to the left. Usually the set $\Omega = [0,\, \tau]$ or $[0,\, \tau)$. Informally, information up to time t consists of all the states the process has visited and the times of the transitions (events). This is made mathematically precise as a σ - algebra, F_t, called also history at time t, consisting of the development of the process in the interval $[0,\, t]$. As the time increases, F_t is an increasing sequence of σ - algebras, a so called filtration. The transition probabilities are defined as:
$P_{hj}(s,\, t) = P(X(t) = j | X(s) = h, F_{s-})$ for $h,\, j \in S,\, s, t \in \Omega,\, s \leq t$
Usually, the multi-state model is defined using its transition (hazard) rates defined as:
$\alpha_{hj}(t) = \lim_{\Delta t \to 0} (\Delta t)^{-1} P(X(t + \Delta t) = j | X(t) = h, F_{t-})$
If $\alpha_{hj}(t)$ only depends on the history via the state $h = X(t)$, the process is called a Markov process. The process is memoryless, in the sense that only the current state is relevant in specifying the transition rates. The rates themselves are allowed to depend on the time t since the beginning of the study. The Markov property is often expressed in the following form: the past and the future of the process are conditionally independent, given the present. In some situations however, there may be dependence on the time since entry to a state. In these cases, we can allow the hazard rate to depend on the time d spent in the current state:

$$\alpha_{hj}(t, d) = \lim_{\Delta t \to 0} (\Delta t)^{-1} P(X(t + \Delta t) = j | X(t) = h, \text{state}\, h \,\text{entered at time}\, t - d)$$

This model is called the extended Markov model, or the modulated Markov model.
If the current state and the time spent in it determine the hazard rate, we have a semi-Markov process, and can write, $\alpha_{hj}(t,d) = \alpha_{hj}(d)$. Another way to express this property is: conditional on the sequence of states, the transition times are independent.
A state $h \in S$ is called absorbing if for all $t \in \Omega, j \in S, j \neq h, \alpha_{hj}(t) = 0$; otherwise h is transient. In our applications there is one absorbing state - death, and subjects are followed until they reach this state, or they are censored in any other state. Next we will describe the illness-death model, where we can consider different types of dependence.

2.3.2 Counting Processes, Likelihood calculations

In this section we will introduce the counting process formulation for time to failure analysis. This approach was introduced in the 1970's by Aalen and has been widely used since then. Assume that the process $X_i(t)$ is observed over the interval $[0,\ \tau_i]$, where τ_i is fixed. Random right censoring and left truncation are considered later. All the information about the process can be derived from the initial state X_0 and the functions

$$N_{hj}^i(t) = (\#\ \text{direct transitions}\, h \longrightarrow j\, \text{in}\, [0,t]\, \text{for}\, subject\, i),$$

where $h, j \in S$. These functions are right continuous and piecewise constant, with jumps at the times of transitions from state h to state j.
For the process i, we will observe $N_{hj}^i(\tau_i)$ such transitions:

$$0 < T_{hj}^{i1} < \cdots < T_{hj}^{iN_{hj}^i} \leq \tau_i,$$

Denote

$$N_{hj}(t) = \sum_{i=1}^{n} N_{hj}^i(t) \text{ and } Y_h^i(t) = I\{X_i(t-) = h\}$$

Note that $Y_h^i(t)$ is an indicator for subject i being in state h, just prior to time t.
The notation $X_i(t-)$ refers to the limit from the left. We have assumed $X(t)$ to be right continuous, with limits from the left. Summing over all individuals, we have (assuming state h is not an absorbing state)

$$Y_h(t) = \sum_{i=1}^{n} Y_h^i(t)\ \ (\ \#\ \text{of individuals 'at risk' in state}\ h\ \text{in time}\ t-\ )$$

Note that for $t > \tau_i$, $N_{hj}(t) = N_{hj}(\tau_i)$ and $Y_h(t) = 0$, so these functions can be considered defined on $(0, \infty)$. The likelihood for the observed process is:

$$\prod_{i=1}^{n} \prod_{h \neq j} exp\left(- \int_0^{\tau_i} \alpha_{hj}^i(t) Y_h^i(t) dt \right) \left\{ \prod_{k=1}^{N_{hj}^i(\tau_i)} \alpha_{hj}^i(T_{hj}^{ik}) \right\} \tag{2.3}$$

This likelihood is somewhat different than the one usually seen in Survival analysis books. If the ith subject is censored in state h, and this is his first visit to the state, then N_{hj} is zero for all states $j \neq h$, and the last term in (2.3) would be missing. In the models we will consider, there will be at most one possible transition $h \to j$, in fact there can be only one visit to every state (e.g. illness-death model).

The derivation above was obtained for observation of subject i up to time τ_i, which is assumed fixed. This is the case with the data set we have - The Framingham Heart Study. In general, the formulas can be extended to two cases of incomplete observation: delayed entry-where individual i enters at some time V_i; and right censoring, where subject i is lost for follow up at time U_i. Both V_i and U_i may be random, but can depend either on the previous history of the process or independent of it. The only correction in the above formulas is the definition of the "at risk" indicator, now: $Y_h^i(t) = I\{X_i(t-) = h,\ V_i < t \leq U_i\}$

2.4 Two examples

Next, we will outline two examples of multi-state models: one is a parametric, the other is non-parametric.

2.4.1 Parametric Hazard rates: constant and piecewise constant

In this model we will assume the hazard for going from state h to state j is constant and the same for all subjects - $\alpha_{hj}^i(t) = \alpha_{hj}$. Let $N_{hj}(T) = N_{hj}$, where $T = max(\tau_1, \ldots, \tau_n)$. Then the likelihood has the form:

$$\prod_{h} \prod_{h \neq j} \alpha_{hj}^{N_{hj}} \, exp\left(-\alpha_{hj} S_{hj}\right)$$

where we assume $0^0 = 1$ and

$$S_h = \sum_i \int_0^T Y_h^i(t) dt$$

The maximum likelihood estimates of α_{hj} are:

$$\hat{\alpha}_{hj} = \frac{N_{hj}}{S_h}, \text{which is the classical occurrence/exposure rate.}$$

An easy generalization is to allow piecewise constant hazard rates: we assume the hazards $\alpha^l_{hj}(t)$ are constant over the interval $(t_{l-1},\, t_l]$. Similarly the maximum likelihood estimates are: $\hat{\alpha}^{(l)}_{hj} = \dfrac{N^{(l)}_{hj}}{S^{(l)}_h}$ where $N^{(l)}_{hj} = N_{hj}(t_l) - N_{hj}(t_{l-1})$ and

$$S_h = \sum_i \int_{t_{l-1}}^{t_l} Y^i_h(t)dt$$

Asymptotic inference can be obtained from the observed information:

$$-D^2 log\, L = \frac{N^{(l)}_{hj}}{(\alpha^{(l)}_{hj})^2}$$

from which we can obtain

$$Var(\hat{\alpha}^{(l)}_{hj}) \sim \frac{(\alpha^{(l)}_{hj})^2}{N^{(l)}_{hj}} \sim \frac{N^{(l)}_{hj}}{(S^{(l)}_h)^2}$$

2.4.2 Freely varying (non-parametric) hazard rates

Assume that the transition rates are the same for all subjects, but vary freely with time $\alpha^i_{hj}(t) = \alpha_{hj}(t)$. We assume the individuals come from a homogeneous population and there are no covariates. Instead of estimating the hazard rates, we can estimate the cumulative hazard:

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u)du$$

using the Nelson-Aalen estimator.

$$\hat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{Y_h(u)} dN_{hj}(u) = \sum_i \sum_{k\,:0<T^{ik}_{hj}<t} \frac{1}{Y_h(T^{ik}_{hj})} \tag{2.4}$$

where $J_h(u) = I\left\{Y_h(u) > 0\right\}$, with variance estimators:

$$\hat{\sigma}^2\left(\hat{A}_{hj}(t)\right) = \int_0^t \frac{J_h(u)}{Y_h(u)^2} dN_{hj}(u) = \sum_i \sum_{k\,:0<T^{ik}_{hj}<t} \frac{1}{Y_h(T^{ik}_{hj})^2}$$

The asymptotic properties of these estimators are derived using techniques based on stochastic integrals and martingales. The matrix of the transition probabilities $\boldsymbol{P}(s,\, t) =$

$(P_{hj}(s,\,t))$ can be calculated as a product integral, which is explained in the next section. Formula (2.6) can be used to obtain an estimator for $\hat{\boldsymbol{P}}(s,\,t)$ by plugging the matrix of the Nelson-Aalen estimators $(\hat{A}_{hj}(t))$ into the formula

$$\hat{\boldsymbol{P}}(s,\,t) = \prod_{s}^{t}(\boldsymbol{I} + \hat{\boldsymbol{A}}(du)) \tag{2.5}$$

This estimator is known as the Aalen-Johansen estimator [16], and for the case of the two-state mortality model , $P_{00}(0,t)$ reduces to the Kaplan-Meier estimator, since $\hat{A}_{00}(t)$ is a scalar step function

$$\hat{A}_{00}(t) = -\hat{A}_{01}(t) = -\sum_{t_j \le t}\frac{d_j}{r_j}$$

and

$$\hat{S}(t) = \prod_{t_j \le t}\left(1 - \frac{d_j}{r_j}\right)$$

where r_j is the number of people at risk just before time t_j, and d_j denotes the number of observed failures at t_j.

The product integral

Multi-state models are usually defined through their intensity functions. Often, the interest is in estimating transition probabilities The matrix of the transition probabilities $\boldsymbol{P}(s,\,t) = (P_{hj}(s,\,t))$ and can be obtained using a (matrix) product integral. Let $\boldsymbol{I}$ be an identity matrix (pxp) and $\boldsymbol{G}$ be a matrix valued function. The product integral is defined as:

$$\Pi_0^t(\boldsymbol{I} + \boldsymbol{G}(ds)) = \lim_{max\,|t_\nu - t_{\nu-1}|\to 0}\Pi(\boldsymbol{I} + \boldsymbol{G}(t_\nu) - \boldsymbol{G}(t_{\nu-1}))$$

where $0 = t_0 < t_1 < \ldots\, t_n = t$ is a partition of $[0,\,t]$
In the case of G - scalar and a continuous function, we have:

$$\prod_{0}^{t}(1 + G(ds)) = e^{G(t)-G(0)}$$

If G is a scalar step function with jumps $d_1, \ldots, d_K$ in the interval $[0,\,t]$, we have

$$\prod_{0}^{t}(1 + G(ds)) = \prod_{k=1}^{K}(1 + d_k)$$

Define $\alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t)$ and the intensity matrix $\boldsymbol{A}(t) = (A_{hj}(t))$, where $A_{hj}(t) = \int_0^t \alpha_{hj}(u)du$. The matrix of the transition probabilities $\boldsymbol{P}(s, t) = (P_{hj}(s, t))$ is given by

$$\boldsymbol{P}(s, t) = \prod_s^t (\boldsymbol{I} + \boldsymbol{A}(du)) \tag{2.6}$$

As an illustration of this formula, consider the case of the exponential distribution for the mortality model, where state 0 is alive, and state 1 is dead. The hazard $\alpha_{01} = \lambda$, $\alpha_{00} = -\lambda$ and the other two intensities $(\alpha_{10}, \alpha_{11})$ are zero. Thus, $A_{00}(t) = -\lambda t$ and $P_{00}(0, t) = \prod_0^t (I + A_{00}(du)) = exp(A_{00}(t)) = exp(-\lambda t)$, since $A_{00}(t)$ is a continuous scalar function. Note that a shorter way to calculate P_{00} is $P_{00}(0, t) = S(t) = P(T > t) = e^{(-\lambda t)}$, for the exponential function.

2.5 Parametric models - Flowgraph model

I will consider the flowgraph model as an illustration for the parametric approach to multi-state models.

2.5.1 Introduction

The main features of the Flowgraph models are:
- parametric assumptions for the waiting time in each state.
- primarily used for data analysis of semi-Markov processes.
- they model the total time to failure, however the hazard can be estimated as well

The flowgraph model, in the form used in the literature, does not involve covariates. On the other hand, it does not assume the proportionality of the hazards as in the Cox model and can be used for multi-state models with very complicated state structures. Flowgraph models have been used for modeling survival data for cancer patients and in reliability. They require the use of a symbolic algebra package, such as Maple, because of the use of approximation techniques (saddlepoint approximation) which are very tedious. We will illustrate the use of flowgraph models for the illness-death model for a fairly homogeneous population- males 50 to 62 years old at baseline of the Framingham Heart Study. As described above, state 0 is healthy state, state 1 is CHD, and state 2 is death.

2.5.2 Censored histograms

Censored histograms are used as a rough guide for choosing a parametric model for a given transition.

Let Y_{ij} denote the waiting time for the transition from state i to state j. We assume the Y_{ij} are independent of each other. This corresponds to a Semi-Markov process. Recall that for Semi-Markov processes, conditional on the sequence of states, the individual waiting times are independent. The Flowgraph method assumes parametric distributions for each waiting time. The choice of the distributions is made using so called censored histograms, which are defined below.

Focusing on each waiting time separately, we can estimate $P(Y_{ij} > t)$ using a Kaplan - Meier estimator. If we are interested in the waiting time Y_{01}, we consider failure as CHD and death before reaching CHD as censoring.

We have

$$P(\tilde{t}_j < Y_{01} \leq \tilde{t}_{j+1}) = S(\tilde{t}_j) - S(\tilde{t}_{j+1}) \approx \hat{S}(\tilde{t}_j) - \hat{S}(\tilde{t}_{j+1})$$

where $\hat{S}(t_j)$ is the Kaplan - Meier estimator for the transition $0 \to 1$.

To construct a censored data histogram, create a set of K equal-width intervals $\left(\tilde{t}_j, \tilde{t}_{j+1}\right]$, $j = 1, \ldots, K$ and over each of them plot a bar with the estimated probability

$$\int_{\tilde{t}_j}^{\tilde{t}_{j+1}} f_{01}(x)dx \approx \hat{S}(\tilde{t}_j) - \hat{S}(\tilde{t}_{j+1})$$

where $f_{01}(x)$ is the density function for the waiting time Y_{01}

The Framingham Heart Study will be introduced in Chapter 3. In this particular data set, we have the death times recorded at every exam, i.e. if a subject dies between exam 5 and 6, his failure time is recorded as 6. As an example I will use the oldest 25% in the male group, that is men older than 49 at baseline. We will use them since they are a more homogeneous group. There are 682 men in this age group and after removing the subjects with CHD recorded at baseline, we are left with 669 men. The censored histogram for this group are presented below.

Based on the censored histograms, we can assume Gamma densities for the transitions $0 \to 1$ and $0 \to 2$, figures 2.2 and 2.3. For the transition $1 \to 2$, we will assume Exponential density, figure 2.4.

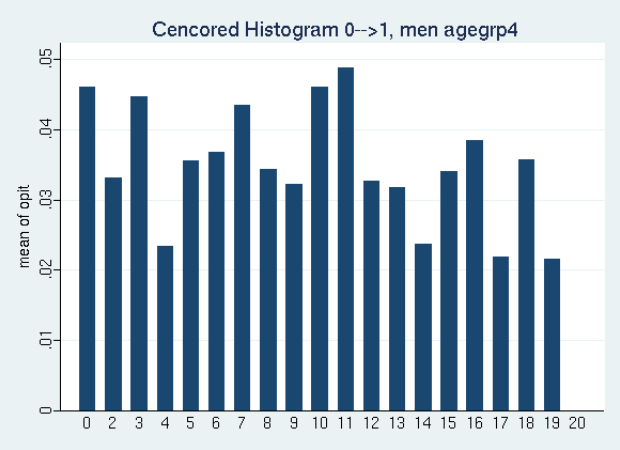

Figure 2.2: Censored histogram, transition from healthy to CHD

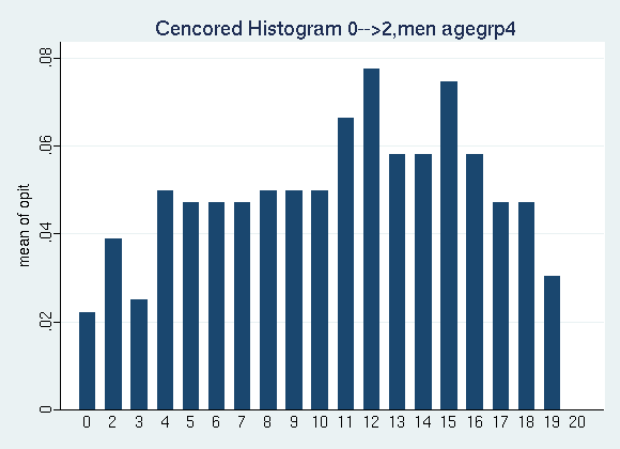

Figure 2.3: Censored histogram, transition from healthy to death

2.5.3 Density function for the total waiting time

Denote Y the total waiting time until reaching state 2. It is $Y = Y_{01} + Y_{12}$ in the case a subject visits state 1 first, and $Y = Y_{02}$ for subjects going directly to state 2. The essence of the flowgraph approach is to estimate the density function for the total waiting time Y. It has two steps, first to obtain the moment generating function (MGF) for Y and second to estimate the density of Y, corresponding to the calculated MGF in step 1. I will outline the

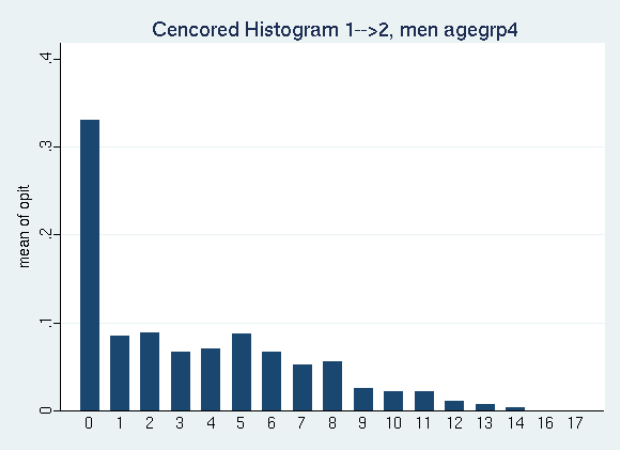

Figure 2.4: Censored histogram, transition from CHD to death

two steps separately.

Calculating the MGF of the total waiting time

The moment generating function (MGF) for a random variable X is defined as

$$M_X(s) = E(exp(sX))$$

provided the expectation exists for s in an open neighborhood of 0.
Let X and Y be two independent random variables and $Z = X + Y$ be their sum. Then, we have $M_Z(s) = M_X(s)M_Y(s)$, the MGF of Z is the product of the MGF of X and Y.
There is a general procedure to obtain the MGF for the total waiting time for any flowgraph. This procedure is based on Mason's rule [17], which was originally developed in the context of graph theory for solving systems of linear equations. For our example in the 3-state illness-death model, we don't need the general rule. Recall the waiting times Y_{ij} are assumed to be independent and $p_{01} = 1 - p_{02} = P(Y_{01} < Y_{02})$. The total waiting time is $Y = Y_{01} + Y_{12}$ with probability p_{01} and $Y = Y_{02}$ with probability $p_{02} = 1 - p_{01}$. Therefore we have

$$M_Y(s) = p_{01}M_{01}(s)M_{12}(s) + p_{02}M_{02}(s)$$

where $M_{ij}(s)$ is the MGF for the waiting time Y_{ij}.

Inverting the MGF of the total waiting time

The flowgraph model gives us the MGF of the waiting time distribution of interest. Our interest is in computing the Bayes predictive density of the total waiting time and this requires converting the MGF to a density function. In special cases, such as convolutions of exponentials, we might be able to find the density of Y by recognizing the MGF, or by using an inverse Laplace transform. In practice, the general approach is to do it numerically using saddlepoint approximation.

Theorem [Saddlepoint Approximation]

Let $K(s) = log[M(s)]$ be the cumulant generating function (CGF). Let c_1 and c_2 be constants such that $c_1 < 0 < c_2$. Suppose $M(s)$ exists for $s \in (c_1, c_2)$, an open neighborhood of zero. Then the saddlepoint approximation for the density of T is

$$\tilde{f}_T(t) = [2\pi K''(\hat{s})]^{-0.5} exp[K(\hat{s}) - \hat{s}t], \tag{2.7}$$

where $K''(s) = d^2K(s)/ds^2$ and the value of $\hat{s}$ is estimated from the so called saddlepoint equation

$$K'(\hat{s}) = t \tag{2.8}$$

There are three main derivations of the saddlepoint theorem, the original is due to Daniels [18], a simpler derivation is due to Barnoff-Nielsen and Cox [19] and the third is due to Barnoff-Nielsen [20]. For most problems involving flowgraphs, $\hat{s}$ is a complicated implicit function of both t and the parameters of the distribution. This is overcome by the use of a symbolic algebra package, e.g. Maple. The constants c_1 and c_2 are found numerically, e.g. the upper bound is the smallest positive root of the saddlepoint equation (2.8). Next, I will illustrate the theorem with an example for a Gamma random variable.

Saddlepoint approximation for Gamma random variable

A Gamma random variable $T \sim Gamma(\alpha, \beta)$ with mean α/β has MGF

$$M(s) = (\frac{\beta}{\beta - \alpha})^{\alpha}, \quad s < \beta$$

The cumulant generating function (CGF) was defined as $K(s) = log[M(s)]$.
In our case, we have $K(s) = \alpha[log(\beta) - log(\beta - s)]$ for $s < \beta$ and

$$K'(s) = \frac{\alpha}{\beta - s} \tag{2.9}$$

$$K''(s) = \frac{\alpha}{(\beta - s)^2} \tag{2.10}$$

The saddlepoint equation (2.8) becomes

$$\frac{\alpha}{\beta - s} = t \tag{2.11}$$

and is solved by

$$\hat{s} = \beta - \frac{\alpha}{t} \tag{2.12}$$

which gives us $K''(\hat{s}) = \frac{t^2}{\alpha}$ and $K(\hat{s}) = \alpha log(\beta) - log(\frac{\alpha}{t})$.
Evaluating the quantities in (2.8) and (2.7), we have

$$\tilde{f}_T(t) = \left(\frac{2\pi t}{\alpha}\right)^{-0.5} exp\left\{\alpha log\left(\frac{\beta t}{\alpha}\right) - t\beta + \alpha\right\}, \ for\ t > 0 \tag{2.13}$$

We can rewrite (2.13) to give

$$\tilde{f}_T(t) = \left(\frac{1}{\sqrt{2\pi}} \frac{\exp(\alpha)}{\alpha^{\alpha - 1/2}}\right) \beta^\alpha t^{\alpha-1} \exp(-\beta t), \ for\ t > 0 \tag{2.14}$$

which is the density function of a Gamma variable up to a constant. The constant is Stirling's approximation to the Gamma function

$$\Gamma(\alpha) \approx \sqrt{2\pi}\alpha^{\alpha - 1/2} e^{-\alpha}$$

In practice, the saddlepoint density is normalized to integrate to 1. The normalizing constant is $\int_0^\infty \tilde{f}_T(t)dt$ and is calculated numerically.

2.5.4 Likelihood calculations

Recall that in the classical survival analysis we have two states, 0 and 1, and we are interested in the waiting time T for the transition $0 \to 1$. Let $f_T(t|\theta)$ and $F_T(t|\theta)$ be the density and the cumulative distribution function (CDF) of T and θ be a vector of parameters. Suppose n subjects are followed and n_1 failures are observed and n_2 censoring times, $n = n_1 + n_2$. Denote the failure times $x_1, \ldots, x_{n_1}$ and the censoring times $x_1^*, \ldots, x_{n_2}^*$. Then the likelihood is

$$L(\theta|data) = \prod_{i=1}^{n_1} f(x_i|\theta) \prod_{j=1}^{n_2} (1 - F(x_j^*|\theta))$$

We will illustrate the likelihood calculations for the illness-death model, and will use data for the oldest male group, age group 4. To derive the likelihood function for this model, let

f_{gh}, F_{gh} be the density and CDF of the waiting time in state g until transition to state h. Based on the censored histograms, we made the following parametric assumptions:

$0 \to 1$, Gamma(α_1, β_1), mean α_1/β_1

$0 \to 2$, Gamma(α_2, β_2), mean α_2/β_2

$1 \to 2$, Exponential(γ), mean $1/\gamma$

The corresponding densities are:

$$f_{01}(t|\alpha_1, \beta_1) = \frac{1}{\Gamma(\alpha_1)} \beta_1^{\alpha_2} t^{\alpha_1 - 1} exp(-\beta_1 t),\ t > 0$$

$$f_{02}(t|\alpha_2, \beta_2) = \frac{1}{\Gamma(\alpha_2)} \beta_2^{\alpha_2} t^{\alpha_2 - 1} exp(-\beta_2 t),\ t > 0$$

$$f_{12}(t|\gamma) = \gamma exp(-\gamma t),\ t > 0$$

Let x_{igh} be the ith uncensored transition from state g to state h, and let x^*_{jg} be the jth censored observation in state g. For example, suppose subject i moves to state 1 at exam 5, and to state 2 at exam 15. Then the subject i contributes data $x_{i01} = 5$ and $x_{i12} = 10$.

We have 669 men in age group 4. State 1 (CHD) was visited by 308 men, out of them 300 reached state 2 and 8 were censored in state 1. There were 341 direct transitions to state 2, and 20 men were censored in state 0. The likelihood contribution from the observations in state 1 is

$$L_1(\alpha_1, \beta_1|\mathcal{D}) = \prod_{i=1}^{300} f_{12}(x_{i12}|\alpha_1, \beta_1) \prod_{j=1}^{8} (1 - F(x^*_{j12}|\alpha_1, \beta_1))$$

where $\mathcal{D}$ denotes the observed data. The contribution from the uncensored observations in state 0 is

$$L_{(0,uncen)}(\alpha_1, \beta_1, \alpha_2, \beta_2, p_{01}|\mathcal{D}) = \prod_{i=1}^{308} p_{01} f_{01}(x_{i01}|\alpha_1, \beta_1) \prod_{i=1}^{341} p_{02} f_{02}(x_{i01}|\alpha_2, \beta_2)$$

and $p_{01} = 1 - p_{02}$ is the probability of moving to state 1, before moving to state 2. The contribution from the censored observations in state 0 is

$$L_{(0,cen)}(\alpha_1, \beta_1, \alpha_2, \beta_2, p_{01}|\mathcal{D}) = \prod_{j=1}^{20} \left\{1 - \left[p_{01} F_{01}(x^*_{j0}) + p_{02} F_{02}(x^*_{j0})\right]\right\}$$

Finally, the total likelihood is the product

$$L(\alpha_1, \beta_1, \alpha_2, \beta_2, \gamma, p_{01}|\mathcal{D}) = L_1 L_{(0,uncen)} L_{(0,cen)}$$

A symbolic algebra package, such as Maple can do the calculations.

2.5.5 Bayesian predictive density

In this section we will describe how to obtain the Bayesian predictive density, survival and hazard functions for the total waiting time in a flowgraph model. Bayesian analysis formally incorporates subjective information about a problem into the analysis via the priors for the parameters. When there is a lack of prior information, one can use non-informative priors. These priors are also called vague or flat priors. With flowgraphs usually independent, non-informative priors are assumed.

The Bayes predictive density of a future failure time T is

$$\begin{aligned} f_T(t|\mathcal{D}) &= \int f_T(t,\theta|\mathcal{D})d\theta \\ &= \int f_T(t|\theta\mathcal{D})\pi(\theta|\mathcal{D})d\theta \\ &= \int f_T(t|\theta)\pi(\theta|\mathcal{D})d\theta \\ &\equiv E_{\theta|\mathcal{D}}\left[f_T(t|\theta)\right] \end{aligned} \tag{2.15}$$

where $\pi(\theta|\mathcal{D})$ is the posterior distribution of θ as defined below.

Definition:[Bayes Theorem]

The posterior distribution of θ, given the data $\mathcal{D}$ is defined by

$$\pi(\theta|\mathcal{D}) \propto L(\theta|\mathcal{D})\pi(\theta) \tag{2.16}$$

where $\pi(\theta)$ is the prior for θ and $L(\theta|\mathcal{D})$ is the likelihood function. We can numerically integrate (2.15) for a given t using Monte Carlo sampling. This involves generating a sample $\theta_1, \ldots \theta_m$ from the posterior $\pi(\theta|\mathcal{D})$ and calculating

$$\hat{f}_T(t|\mathcal{D}) = \frac{\sum_{j=1}^{m} \tilde{f}_T(t|\theta_j)}{m} \tag{2.17}$$

where , $\hat{f}_T(t|\theta)$ is the estimated density from the saddlepoint approximation.

A similar method has been presented in Welton and Ades [21]. They proposed a model, where the transitions $i \rightarrow j$ are assumed with constant intensities α_j, and a Bayesian approach is used to find the posterior densities for these rates. They consider the situation with fully observed data (we know all transitions and the time they occurred) as well as a situation where some transitions may have not been observed. However, as they pointed out, the assumption of constant hazard rates is unrealistic and some variations are discussed.

2.6 Life Table calculations

The classical life table is designed to deal with two states - alive and death. The time is age for the individuals, and the time span is divided into age periods. For each age interval the number of people who die, withdraw or enter the study is recorded. The literature on life tables is vast and different versions of tables have been developed. Chiang ([22]) has an extensive treatment of life tables. The quantities in which we are interested are the proportion of the survivors at age x and the life expectancy (also denoted LE) at age x. The construction of a life table will be outlined briefly, following Chiang ([22]). For simplicity, we will illustrate the case for a life table without entries and withdrawal during the follow up.

For each age interval $[x, x+1)$ the following quantities are recorded:

l_x - number of alive at age x,

$l_{x+1} = l_x - d_x$

d_x - number of people dying in the interval $[x, x+1)$.

$\hat{q}_x$ - proportion of people dying in the interval $[x, x+1)$,

$\hat{q}_x = \frac{d_x}{l_x}$

L_x - number of years lived in the interval $[x, x+1)$,

$L_x = l_x - \frac{1}{2}d_x$ assuming on average people survived the first half of the interval.

T_x - total number of years lived beyond age x,

$T_x = L_x + T_{x+1}$

$\hat{e}_x$ - life expectancy at age x,

$\hat{e}_x = \frac{T_x}{l_x}$

The life expectancy e_x summarizes the mortality experience for a subject beyond age x. The quantity $p_x = 1 - q_x$ is the probability of surviving in the age interval $[x, x+1)$, given he/she has survived till age x. The probability of surviving until age x, $S(x)$ can be expressed as the product

$S(x) = p_0 p_1 \dots p_x$

The life table 2.1 is adjusted from Chiang, [22] and it describes the California population in 1960. The initial number l_x - number of alive at age x can be arbitrary radix, here chosen to be 100,000. Table 2.1 shows that of every 100,000 persons born in alive will 97,496 will survive their second birthday (at age 3), provided that experience the same mortality as the

California population in 1960.

Table 2.1: Life Table for the total California population, 1960

Age	l_x	d_x	$\hat{q}_x$	L_x	T_x	$\hat{e}_x$
0-1	100,000	2378	.2378	97,860	7,058,410	70.58
1-2	97,622	146	.00150	97,539	6,960,550	71.30
2-3	97,476	95	.00097	97,424	6,863,011	70.41
3-4	97,381	77	.00079	97,340	6,765,587	69.48
4-5	97,304	58	.00060	97,274	6,668,247	68.53

2.6.1 Multi-State Life Tables

Life tables can be generalized considering more than 2 states. Peeters ([10]) presents life tables with states CVD, CHD or CVA, one at a time, added to the healthy state and the death state. In this case, all the formula above would be state specific. For example $d_{ij}(x)$ is the number of transitions from state i to state j in the interval $[x, x+1)$. Focusing on one transition at a time, the probability of developing the disease can be estimated. If we focus on the states $i = healthy$ and $j = CHD$, we need to treat CHD as a failure, and death as censoring. Petters ([10]) has estimated the probabilities of developing different types of CVD, using a Multi-state life table based on the cohort from the Framingham Heart Study. Some of the estimated probabilities are presented in the table below:

Table 2.2: Lifetime risk of developing CVD for individuals free of CVD at age 40

Probability of developing a disease			
Males			
CVD	*Acute M.I.*	*Stroke.*	*CHF*
.67	.32	.16	.18
Females			
CVD	*Acute M.I.*	*Stroke.*	*CHF*
.59	.17	.21	.19

2.6.2 State specific life expectancies

Life expectancies can be calculated from the estimated survival function specific for a particular state. In this way the life expectancies for a healthy subject can be compared with the estimated time free of a particular disease. The results from Peeters ([10]) are presented in Table (2.3) below.

Table 2.3: Life expectancy (LE) and residual LE free of disease

LE and residual LE free of disease						
Males		LE free of history of				
Age	*LE*	*CVD*	*CHD*	*M.I.*	*CVA*	*CHF*
50	26.2	19.9	21.5	23.3	25.2	25.5
70	12.0	7.38	8.69	9.94	11.2	11.4
Females		LE free of history of				
Age	*LE*	*CVD*	*CHD*	*M.I.*	*CVA*	*CHF*
50	32.1	26.4	28.4	30.9	30.9	31.2
70	16.0	11.3	13.0	15.0	14.9	15.2

At age 50, the life expectancy for a male is 26.2 years and the life expectancy free of CVD is 19.9. This leaves 6.3 years of the expected residual life spent with CVD. For women at age 50, 5.7 years are estimated to be spent with CVD. Men have higher estimates for expected years lived with CVD at both age 50 and 70. On the other hand 50 year old females will live on average 0.29 more years with stroke and 0.26 more with CHF compared with males. As indicated by the authors, the greater longevity of females leads to higher burden of some of the diseases.

The multi-state life table method is designed primarily for situations in which actual failure and censoring times are unavailable and only the total numbers are given. They are appropriate for a homogeneous population, since the effects of covariates are not considered.

CHAPTER 3

Three-state Models

3.1 The illness-death model

We introduced the mortality and the competing risk models in Chapter 2. These models have been used extensively for modeling the risk of binary outcomes. As the population continues to age however, the process of moving from health to death becomes more complex with larger and larger proportions of the population living with prevalent chronic diseases. This Chapter deals with the next step in understanding this process of movement between states; it focuses on the 3 state health, illness, and death model.

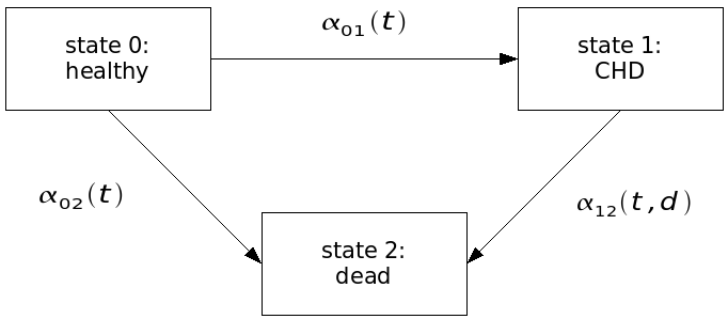

Figure 3.1: Illness-death model

Figure 3.1 presents the three state model commonly referred to as the illness-death

model. The 3 states, healthy, illness and death, are denoted state 0, state 1 and state 2 respectively. The illness-death model has been widely used to compare mortality before and after a particular event such as heart transplant (Hougaard [23]) or bone marrow transplant (Aalen [13]). This model was discussed in early papers by Fix and Neyman [24] and Sverdup [25]. There has been less use of the model for considering the interrelationship of health, chronic diseases such are cardiovascular or cerebrovascular disease, and death.

There are two equivalent methods for applying the multi-state approach to a particular problem:

- Estimate the transition (hazard) rates for moving between different states or equivalently,
- Combine the transition rates to obtain transition probabilities.

In this chapter, we will use both approaches.

In sections 3.4.2 and 3.4.1 we will examine methods for estimating the transition rates $0 \rightarrow 1$ and $0 \rightarrow 2$. The goal is to compare the two transition rates, before and after CHD. This could be used to evaluate the role CHD plays in mortality.

In section 3.5 we estimate all three hazard rates and combine these estimates (often called a synthesis of a multi-state model), in order to calculate transition probabilities.

Transition probabilities are the conditional probabilities of moving to a particular state given their previous history and can be used for estimating the prognosis of individuals, given their event history (in this case whether they have developed CHD) at a given time.

In order to compare the hazard rates, we have to decide what they would have in common and in what ways they will differ. In this chapter we assumed the two hazard rates (for moving from state 0 to state 2 and for moving from state 1 to state 2) are proportional, i.e. they have the same baseline hazard up to a constant of proportionality that is a function of their characteristics. In section 3.4.2 the two transition rates share the same covariate effects (coefficients) and the effect of CHD is modeled as a multiplicative factor increasing the hazard after a person enters the CHD state.

In section 3.5, the effect of CHD is modeled as the covariates are considered to have different effects before and after CHD. Next, we will introduce some notation and definitions.

Denote t_1, the time of entering state 1 (for the subjects who enter it) and $d = t - t_1$ time since entering state 1. We have 3 possibilities for modeling the hazard rate α_{12} for moving from state 1 to state 2 depending on the assumption we make:

- If we assume $\alpha_{12}(t)$ does not depend on d, we have a Markov model.
- If we assume the hazard rate $\alpha_{12}(t, d)$ depends on the time in state 1, the model is an extended-Markov model. (Note that we can write $\alpha_{12}(t, d)$ or $\alpha_{12}(t, t_1)$ because of the linear relation $d = t - t_1$.)
- In the special case $\alpha_{12}(d)$ is a function of d, the time spent in state 1 alone, the model is called semi-Markov.

3.2 Data

3.2.1 The Framingham Heart Study.

The Framingham Heart Study was designed by The National Heart, Lung and Blood Institute (NHLBI) (formerly known as the National Heart Institute) to identify common factors or characteristics that contribute to cardiovascular disease (CVD). The study began by recruiting an original cohort of 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts in the period 1948-1952. The cohort has been followed since then, with exams every two years. Data on the vital status and the occurrence of CVD was recorded. Statistical analysis of the data has identified several major CVD risk factors, as well as information on the effects of these factors such as blood pressure, blood triglyceride and HDL cholesterol levels, age, gender, smoking etc. An Offspring Cohort was added in 1971, and a Third Generation Cohort began in 2000.

3.2.2 The data files.

The data set we will use contains the exact dates of CHD or death for 4,266 people from the Framingham Heart study. A new procedure for measuring cholesterol was adopted at exam 4 onwards and is considered more reliable. We will use exam 4 as a baseline (follow up time $t = 0$). The other covariates in our analysis were measured at exam 4. The last time of death recorded is at 37.69 years. There were 518 individuals who had experienced at least one event

(CHD, CVA or death) before exam 4, and were excluded from the analysis. We also excluded the individuals with missing at least one of the covariates cholesterol, smoking status, age at exam 4 , diabetes or systolic blood pressure. There were 3,201 individuals (1,520 men and 1,681 women) who were included in the analysis. There were 2,722 deaths and 1,400 CHD times recorded. The data were acquired through a limited public use agreement with the National Heart, Lung, and Blood Institute.

3.3 Exploratory analysis

We introduced the Framingham heart study in section 3.2 and the data set which will be used in this section. The variables considered in our analysis are age at exam 4, diabetes, smoking status, systolic blood pressure, total cholesterol level and gender. Three new covariates will be used as well: z is defined as an indicator for being in state 1, and c_risk is an indicator for just moving to state 1 and d is the time spent with CHD. They are time dependent variables and do not depend on the the baseline covariates, rather they describe the transitions in the 3 state model. All variables are presented in Table 3.1.

Table 3.1: Variables

Categorical Variables	
psm	indicator for smoking
diab	indicator for diabetes
z	indicator for being in state 1 (CHD state)
c_risk	indicator for just entering state 1
Continuous Variables	
age	age at exam 4 in years
spf	systolic blood pressure
chol	total cholesterol level
d	time with CHD in years

We described the illness-death model in section 3.1. It consists of 3 states: healthy state, illness and death, denoted state 0, 1 and 2 correspondingly. Initially every subject was in the healthy state 0 and there were no drop outs from the study. Among the 1,520 men in the group, 767 visited state 1 (CHD),among them 637 moved to state 2 (died). There were 683 males who moved directly to state 2. At the end of the study, 130 men were censored

in state 1 and 70 were censored in state 0.
From the 1,681 women in the study, 633 visited state 1, and from them 468 eventually moved to state 2. There were 934 direct transitions from state 0 to state 2 and 114 women censored in state 0.

The most noticeable differences between the two genders are:

- 27% of the women have both events, versus 42% of the men
- 56% of the women died without CHD, versus 45% of the men

It appears CHD plays a bigger role in mortality in the male group.
For the 637 men and 468 women who died with CHD, the distribution of the time spent with CHD is presented as histograms in figures 3.2 and 3.3.

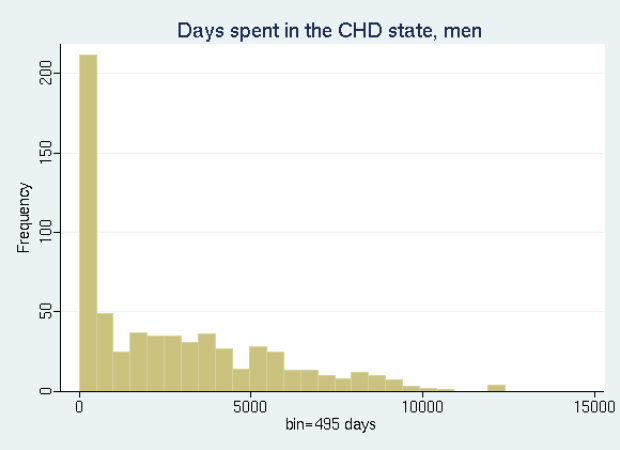

Figure 3.2: Time spent in the CHD state, men

In both cases, there is a large number of people failing very shortly after being diagnosed with CHD. One of the components of CHD is sudden coronary death, which is defined as a death which occurs within 24 hours of the onset of CHD. In the male group, this type of death accounts for 134 cases, which is 10% of the total deaths. Similarly, in the female group, we have 94 deaths (6% of the total deaths). There are two points to be made here. First is that CHD in females includes a higher proportion of Angina Pectoris (AP), which is

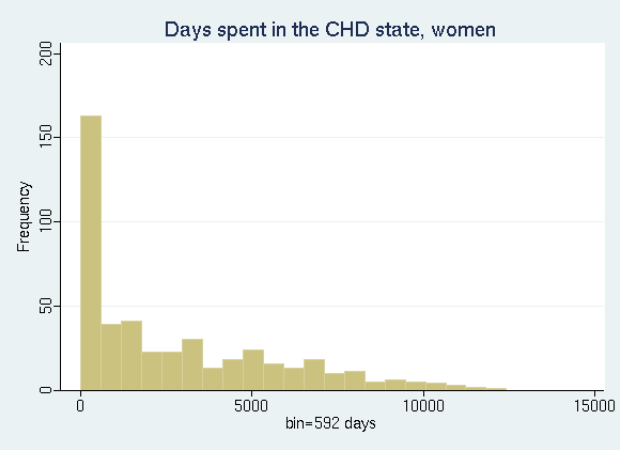

Figure 3.3: Time spent in the CHD state, women

milder form of CHD, compared to Myocardial Infarction (MI). We have 50.98% MI in the male group (23.86% recognized with ECG) versus 39.18% for women (17.38% recognized with ECG). On the other hand men have 30.51% AP, while it accounts for 43.13% of the CHD cases in women.

The other point is that the proportion of sudden deaths has changed in the last decades perhaps due to advances and wider availability of medications and preventive health care. To look for a possible trend we calculated the proportion of sudden deaths (to the total number of deaths) for each year. The results are presented separately for men and women in figures 3.4 and 3.5.

We can observe a significant drop around the 12^{th} year of follow up in the male group. In this group, the average number of sudden deaths per year is 3.37 and the total average number of deaths is 34.73. The proportions fluctuate due to the relatively small number of deaths, when broken into one year periods. The 12^{th} year of follow up corresponds to the early 1970's. Anderson [7] states that the early 1970s is a time "when there was a sharp decrease in CHD mortality rates in US and extensive intervention against risk factors being practiced".

The covariate *c_risk* was defined as an indicator for just moving in state 1. It will be useful in accounting for the failures occurring immediately after moving to state 1, as well as measuring the downward trend for the percentage of sudden deaths.

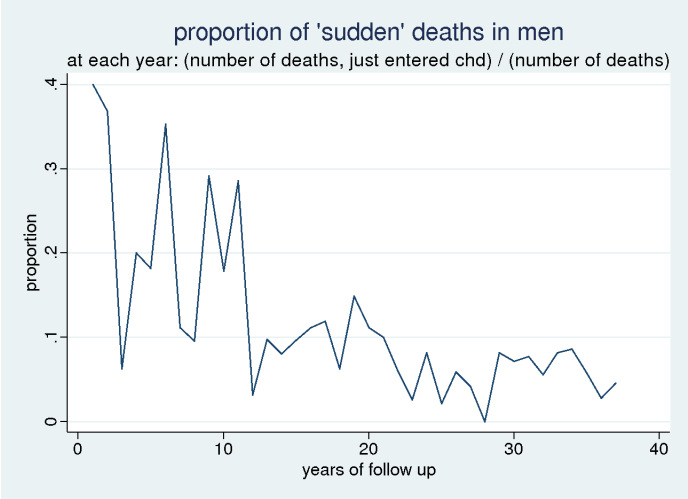

Figure 3.4: Proportion of sudden deaths, men

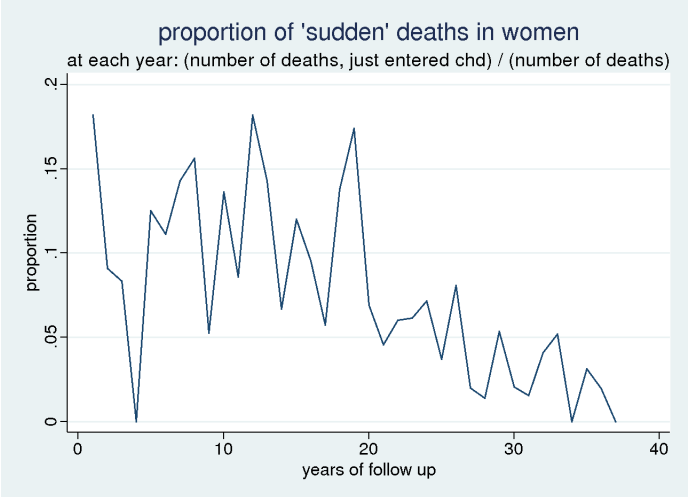

Figure 3.5: Proportion of sudden deaths, women

Time in our analysis is the follow up time. This is the case with most of the epidemiological studies. Using age as the time scale has been proposed in the literature, as well [9].

The Framingham Heart study is an observational study with more than 45 years of follow up. With such a long time span, the use of follow up time presents challenges. For example, if we use the Cox model, the coefficient for age β_{age} has the interpretation of the hazards ratio, for two individuals, one being 1 year older than the other. In other words, the hazard ratio for two men age 40 and 50 at the beginning of the study should remain the same 20 years later, when they are 60 and 70. This assumption may not be satisfied.

Another challenge related with the long follow up period is that the covariates are recorded at baseline. Smoking status for example was recorded in the early 1950's when smoking was common, particularly in men. In the 1960's many people quit smoking and the effect of this variable may be expected to be reduced with time.

These two observation, as well as the analysis that we will perform later, suggest that the effect of some variables may change with time. In order to capture this, we will consider models where the coefficients in the Cox model are not constants, but piecewise constants. To do so, we will split the follow up time into four time intervals:

$$[\tau_0, \tau_1), [\tau_1, \tau_2), [\tau_2, \tau_3) \; and \; [\tau_3, T)$$

where $\tau_0 = 0$ corresponds to the baseline (exam 4), and the other cut off points in years are: $\tau_1 = 8.7$, $\tau_2 = 15.7$, $\tau_3 = 24.7$ and $T = 37.7$ is the end of the study.

The choice of the other split points were made so that we have a split point corresponding to early 1970's and to have an interval in the beginning, where figures 3.4 and 3.5 suggest CHD played a bigger role in mortality.

We will adopt the following convention, for any variable v, we denote:

$$vXj = v\, I[\tau_{(j-1)} \leq t < \tau_j)$$

i.e. the covariate v for the duration of the period $[\tau_{(j-1)}, \tau_j)$. For example, we define the covariate $z(t)$ to be an indicator for having CHD. Thus, $zX1 = 1$ for a subject if at time t he/she is in state 1 and $t < \tau_1$. We have $z = zX1 + zX2 + zX3 + zX4$, and this will allow us to model the effect of CHD with a coefficient that is a step function.

3.4 Multi-state models for the illness-death model

As outlined in Chapter 2, the most general case would require that hazards be modeled separately for each transition. In the Framingham study, it is particularly interesting to compare the hazard rates before and after entering the disease state. We will use models based on the Cox model and the Aalen additive model.

3.4.1 Markov Models

The simplest model is a Markov model that assumes the hazards for the transitions $0 \to 2$ and $1 \to 2$ are proportional, i.e. $\alpha_{02}(t) = c\alpha_{01}(t)$. In other words the hazard of dying changes by a factor the instant a subject gets CHD, regardless of the time this happens or the covariates. Under the Cox proportional model, the hazard rates are:

$$\alpha^i_{02}(t) = \alpha_0(t)exp(\boldsymbol{\beta}'\boldsymbol{X}_i)$$

$$\alpha^i_{12}(t) = \alpha_0(t)exp(\boldsymbol{\beta}'\boldsymbol{X}_i + \beta_z)$$

where $\beta_z = log(c)$ and $\boldsymbol{X_i}$ is the vector of covariates for subject i. The covariate z is an indicator for having CHD and was introduced above. Strictly, it can be defined as follows:

$$z(t) = I[t > u]$$

It is zero before the time u of getting CHD, and one afterwards. For the individuals who do not get CHD, $u = \infty$. This is an example of a Markov model. We fit separate models for males and females. The results are presented in Table 3.2.

The coefficient for age in the female group is modeled as a step function: it is .046 for the first two time periods of the follow up time, and .073 for the third and fourth time periods. The variables $age1half = ageX1 + ageX2$ and $age2half = ageX3 + ageX4$ are used, since we found that if we use the variables $ageXj$ for $j = 1, \ldots, 4$ the first two coefficients are similar, and far apart from the other two. The Likelihood ratio test did not show a better fit for a model with all four coefficients (p-value 0.33). We will comment on this fact later, when we compare it with the extended Markov model.

The coefficient β_z suggest that a woman with CHD has 3.15 times higher hazard of dying in the first three time intervals. The hazard ratio reduces to 1.10 for the fourth time period.

Table 3.2: Markov Models

Markov Model for Females						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\|z\|$	*95 % Conf. Interval*	
$age1half$	.0458423	.0076186	6.02	0.000	.0309101	.0607746
$age2half$	.0735431	.0043828	16.78	0.000	.0649529	.0821332
diab	.7680075	.155011	4.95	0.000	.4641916	1.071823
psm	.3301785	.0576088	5.73	0.000	.2172673	.4430897
spf	.0065403	.0011432	5.72	0.000	.0042997	.008781
z	1.149581	.08949	12.85	0.000	.9741837	1.324978
$zX4$	-1.050464	.1183848	-8.87	0.000	-1.282494	-.8184344
Markov Model for Males						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\|z\|$	*95 % Conf. Interval*	
age	.0757493	.0038013	19.93	0.000	.0682988	.0831999
diab	.4692469	.1469549	3.19	0.001	.1812206	.7572733
psm	.3482076	.0657918	5.29	0.000	.2192581	.477157
spf	.0128758	.0017516	7.35	0.000	.0094428	.0163089
$spfX4$	-.0107079	.0029592	-3.62	0.000	-.0165079	-.004908
chol	-.0026135	.0007346	-3.56	0.000	-.0040533	-.0011736
$zX1$	2.613352	.192252	13.59	0.000	2.236545	2.990159
$zX2$	1.473335	.1446044	10.19	0.000	1.189916	1.756755
$zX3$	1.207545	.1020531	11.83	0.000	1.007525	1.407565
$zX4$	.4848401	.0833764	5.82	0.000	.3214254	.6482549

We compare different models using the Likelihood ratio test, which is described in detail in the Appendix.

The model for males requires the coefficient for the variable z to be modeled as a step function as well. During the first time interval, the proportion of sudden deaths was very high and is reflected in the hazard ratio, estimated as 13.64. The hazard ratio estimated for the fourth interval 1.62. The coefficient for the cholesterol level is negative. We will compare this result with the estimate from the Aalen model. The role of this covariate for getting CHD is examined in section 3.4.4, here we model the hazards of dying.

The coefficient for blood pressure is modeled as a step function, with a separate constant for the fourth interval. The effect is reduced substantially. This may be attributed to the long follow up period and its variability. Even if we assume that the two measures are highly correlated, the availability of effective drugs regulating high blood pressure in the later time

periods inevitably reduced its effect.

Using only an indicator for the CHD, the coefficients reflects both the burden of sudden deaths as well as for deaths which occur after a long time spent with CHD. In order to account for the sudden deaths, we will use an extended Markov model, i.e. we will allow the hazard to depend on the time spent in the CHD state.
A test based on the scaled Shoenfeld residuals was used to test for the proportionality assumption (described in the Appendix). The combined p-values for these test were .06 for females and .14 for males. The variable z has the most contribution for the large test statistics (small p-value) in the model for women.

3.4.2 Extended Markov Models

The models we used in the previous section assumed that having CHD increases the mortality by a factor, which is a constant, i.e. the same for the first day after getting CHD as one year later. As we discussed in section 3.3, the mortality is very high at the beginning, particularly the first day (sudden coronary death). The next step in our analysis is to include time since diagnoses as a covariate. This leads to a general class of Markov models - the so called Extended Markov models. The hazard rates are modeled as follows:

$$\alpha_{12}^{i}(t) = \alpha_0(t) exp(\boldsymbol{\beta}' \boldsymbol{X}_i + \beta_z + f(d))$$

where d is the time spent in state 1 and $f(d)$, $d \geq 0$ is any function and z is the indicator for entering the CHD state introduced above. The function $f(d)$ needs to be defined for $d = 0, 1, \ldots$ which represents the number of years after the diagnoses.

For a small data set, an extended Markov model can be fit with splitting the records at the failure times, that is the record for subject i will be split at any time somebody dies before he/she does. This is not feasible for a data set with more than a thousand observations. As reported by STATA, this will create more that 2,000,000 observations. In order to avoid this, we rounded the death times and the CHD times to the next year. Potentially, this may artificially increase the number of people who has died immediately after getting CHD. However, in our data set, among the 357 people with CHD and death within a year apart, 216 died within a day from the diagnoses. For the rest of this section, we will work with both events measured in years.

We will use two parameters to describe the function, one for the value $f(0)$ and the other for a linear trend $f(d) = kd, d = 1, 2 \ldots$. The value of the function at $d = 0$ will be modeled as a separate covariate c_risk, which is defined to be one only on the year the diagnoses is made for CHD. Therefore, this is a time varying covariate and its value for subject i will be 1 for at most one year. If subject i visits state 1, the covariate c_risk has the value 1 for just that year.

Often it is the case, that if a subject survives a certain amount of time after getting the disease, his/her chances improve. The second coefficient k, used to determine the function $f(d)$ measures the long term effect of the disease. If a subject has spent d years in state 1, the hazards ratio, compared with a person without CHD is $\exp(\beta_z + kd)$. If the coefficient k has negative value, it will indicate the chances of a patient are improving the longer they state in state 1, compared with a person who has spent less time in state 1.

Let d_1 and d_2 be the time spent in state 1 for two subjects with CHD. Their hazards ratio is $\exp(k(d_1 - d_2))$, as long as both of them has survived the first year. One limitation of modeling the function $f(d)$ with only one coefficient for the range $d \geq 0$ is that the hazards ratio depends only on the difference of the time spent in state 1 , i.e. the same for ($d_1 = 2$, $d_2 = 3$) and ($d_1 = 15$, $d_2 = 16$).

Next, the results for males and females are presented in Table 3.3 using the Extended Markov models.

The model for females has coefficient β_{age} which is a step function, with very similar values to the Markov model. The coefficients for diabetes and smoking status are slightly smaller in the Extended Markov models in both genders. The same holds for the coefficients for the indicator variable z. This was expected, since in the Markov model, the coefficient has to account for both the very high rate of sudden coronary deaths, as well as for the increased mortality in general.

The Markov and the extended Markov models have the coefficients for the variable z as step functions. In the both groups, the step function has two different values, one for the first 3 time periods and a noticeably smaller coefficient for the last time period. Note, that for the fourth period, the coefficient for the indicator z is $\beta_z + \beta_{zX4}$. Therefore, the coefficient for the fourth intervals are .27 in the male group and .08 for females.

Table 3.3: Extended Markov Models

Extended Markov Model for Females						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\|z\|$	*95 % Conf. Interval*	
$age1half$	.045257	.0075757	5.97	0.000	.0304088	.0601051
$age2half$	.0701719	.0043682	16.06	0.000	.0616104	.0787334
diab	.6464857	.1561526	4.14	0.000	.3404322	.9525392
psm	.2938729	.0574288	5.12	0.000	.1813145	.4064313
spf	.0063993	.0011363	5.63	0.000	.0041722	.0086264
c_risk	1.651802	.1529767	10.80	0.000	1.351973	1.951631
c_riskX4	-.5922909	.21503	-2.75	0.006	-1.013742	-.1708397
z	.6928349	.1205777	5.75	0.000	.4565069	.9291628
$zX4$	-.6100307	.1418639	-4.30	0.000	-.8880788	-.3319826
d	-.0144801	.0076415	-1.89	0.058	-.0294571	.0004969
Extended Markov Model for Males						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\|z\|$	*95 % Conf. Interval*	
age	.0718955	.0037845	19.00	0.000	.0644781	.0793129
diab	.3909468	.147081	2.66	0.008	.1026734	.6792202
psm	.3279029	.0656417	5.00	0.000	.1992475	.4565582
spf	.0104134	.0016919	6.15	0.000	.0070974	.0137294
$spfX4$	-.0082601	.0029135	-2.84	0.005	-.0139704	-.0025498
chol	-.0021518	.0007303	-2.95	0.003	-.0035832	-.0007203
c_riskX1	2.185154	.2112392	10.34	0.000	1.771133	2.599176
c_riskX2	1.619721	.1890416	8.57	0.000	1.249206	1.990236
c_riskX3	1.16372	.1582187	7.36	0.000	.853617	1.473823
c_riskX4	1.060161	.1508432	7.03	0.000	.764514	1.355808
z	.9098498	.0907009	10.03	0.000	.7320793	1.08762
$zX4$	-.6375623	.1257962	-5.07	0.000	-.8841183	-.3910063

The variable *c_risk* was not included in the Markov models, since it depends on d - the time spent in the CHD state (in this case d=0). This is an indicator variable, for just entering the CHD state. Strictly speaking, the indicator is one if the CHD and death occurred in the same year, since the CHD and death dates were rounded to the beginning of the next calendar year. This variable accounts precisely for the high rate of mortality for people just entering state 1. The figures 3.4 and 3.5 show that the effect of this variable is reducing for the years when the Framingham study took place. The coefficients in our model reflect this downward trend - as in the figures this is more noticeable in the males group.
Cholesterol is a significant variable for males, with a coefficient suggesting beneficial effect.

This was also observed in the Markov model and we commented there for possible reasons. Later, we will see that the effect of cholesterol for the transition from the healthy state to the CHD state is adverse.
So far we discussed the differences between the Markov and the Extended Markov models for both genders. We have observed similar differences/similarities. Next, we will compare the Extended Markov models for the two gender groups.

The effect of CHD to mortality can be seen through the coefficients for the variables indicating transitions in the multi-state model, i.e. the variables z and c_risk. For both of them the effect is stronger in the male group. The effect of entering the CHD state is modeled as a step function with four different values in the male group, while two values are sufficient for females. This indicates that the effect of CHD, particularly sudden deaths reduced more significantly in the male group. Another reason may be that there were more males dying after just entering the CHD state and therefore make it more noticeable (the test for significance of the coefficients has more power).
The effect of the covariate d - time spent in the CHD state was not significant for males and has a negative value for females. This indicates reducing the effect of CHD with the time spent in this state in women. We observed in the exploratory analysis that women spent more time in the CHD state and the effect of CHD to mortality appears to be smaller.

In summary, the effect of CHD to mortality in the male group was modeled as a factor to the hazards ratio - either $\exp(\beta_z + \beta_{c\_risk})$ initially and $exp(\beta_z)$ later on. The mortality is higher at the beginning and stabilizes afterwards.
In the female group the effect of CHD was modeled as a factor: $\exp(\beta_z + \beta_{c\_risk})$ initially and $\exp(\beta_z + d\beta_d)$, which reduces with the time a women stays with CHD.

One difference between the models for males and females, in both the Markov and the Extended Markov models, is that the coefficient for age is modeled as a step function for females. As we have commented in the beginning of the Chapter, this is to be expected, since otherwise we assume that the hazard ratio for two individuals age 40 and 50 will remain the same, even when they are 70 and 80. Therefore, we expect it to increase, and in the model for male a variable $ageX4$ has a borderline p-value 0.07. If we use this model we will obtain the coefficients $\beta_{age} = .069$ and $\beta_{ageX4} = .013$. Therefore, the effect of age will be modeled as a step function, with the value 0.069 for the first three time periods and 0.082 for the last period.

In both models for males, i.e. if we use a model with a constant as a coefficient for age (covariate *age*) or a step function (variables $ageX4$ and *age*) the effect of age changes less for the duration of the follow up. We have a noticeable difference (0.025) in the female group for the first and the second half of follow up. One explanation we can offer is that the effect of CHD is stronger in males, and accounting for it leaves less explanatory power for age. In the female group, particularly since many women live many years in the CHD state, age is more important - loosely speaking, at the end they often die of old age, even if they suffer from CHD. Another reason is that in the first half, fewer women died, so the effect is not so pronounced.

We compare the coefficients for the remaining variables: the effect of smoking and systolic blood pressure is slightly higher in males, the effect of diabetes is higher in women - a result reported by other studies as well.

A test based on the scaled Shoenfeld residuals was used for the proportionality assumption. The test is detailed in the Appendix. The combined p-values for this test were .48 for females and .58 for males.

3.4.3 The Aalen additive model

The Aalen additive model was introduced in Chapter 2. The main difference from the Cox model is that the hazards are modeled as a linear combination of the covariates. The Aalen model allows for time varying coefficients and time varying covariates. The hazard function for the transition $i \to j$ is modeled as:

$$\alpha_{ij}(t, \boldsymbol{X}) = \beta_0(t) + \beta_1(t)X_1(t) + \ldots + \beta_k(t)X_k(t) \tag{3.1}$$

In our case, we have fixed time covariates, but the possibility of having time-varying coefficients is appealing, having discovered in the previous section that the effect of some covariates changes with time.

The regression coefficients $\beta_h(t)$, and possibly also the covariates will depend on which pair of states is considered. Let's assume there are k covariates considered for the transition (i, j) and we observe n individuals. The design matrix $\boldsymbol{Y}_{ij}(t)$ $(n \times k + 1)$ is defined as follows:
There is one row for each subject, if he is still at risk, the row is $(1, X_1(t), \ldots, X_k(t))$, and zero if the individual is not at risk. The multivariate intensity process for the multivariate counting process for the transition (i, j) can be written in the matrix form:

$$\lambda_{ij}(t) = \boldsymbol{Y}_{ij}(t)\boldsymbol{\beta}_{ij}(t)$$

where $\boldsymbol{\beta}_{ij}(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_k(t))'$. We are interested in estimating the cumulative regression coefficient functions $B_h(t) = \int_0^t \beta_h(u)du$. The vector of these functions for the (i, j) transition is denoted by $\boldsymbol{B}_{ij}(t)$. Following Aalen [12] the vector $\boldsymbol{B}_{ij}(t)$ can be estimated by:

$$\hat{\boldsymbol{B}}_{ij}(t) = \int_0^t \boldsymbol{H}_{ij}(s)d\boldsymbol{N}_{ij}(s)$$

where $\boldsymbol{H}_{ij}(t)$ is the generalized inverse of $\boldsymbol{Y}_{ij}(t)$ and $\boldsymbol{N}_{ij}(t)$ is the multivariate counting process, counting for each individual the transitions from state i to state j. The estimation procedure is well defined only for the time when the matrix $\boldsymbol{Y}_{ij}(t)$ has full rank. The matrix $\boldsymbol{H}_{ij}(t)$ is calculated as:

$$\boldsymbol{H}_{ij}(t) = (\boldsymbol{Y}'_{ij}(t)\boldsymbol{Y}_{ij}(t))^{-1}\boldsymbol{Y}'_{ij}(t)$$

As noted in Klein and Moeschberger [26], this is a least-square estimation technique. For comparison, the Cox proportional hazards model relies on likelihood based estimation. The basis for the statistical theory of the estimator $\hat{\boldsymbol{B}}_{ij}(t)$ is the fact that $\hat{\boldsymbol{B}}_{ij}(t) - \boldsymbol{B}_{ij}(t)$ is a martingale. Addreg - a program in R, has been developed by Aalen and Fekjer and is available online, (see [13]) which fits the Aalen model. We have used Addreg to fit the Aalen model for each for the three transitions $0 \to 1$, $0 \to 2$ and $1 \to 2$ separately. This is a major difference from the approach we used with the Cox model- we assumed the two hazard rates $0 \to 2$ and $1 \to 2$ to be proportional. Further, the Aalen model obtains estimates for the cumulative hazard function $\boldsymbol{B}_{ij}(t)$ and not for the coefficient function $\beta_{ij}(t)$. Smoothing estimators have been considered for $\beta_h(t)$, Aalen [27], but this is beyond of the scope for this thesis. The cumulative coefficient function is an integral of the coefficient function and mathematically can be obtained by differentiating. However, the estimate $\hat{B}_{ij}(t)$ is a step function and differentiating is not possible. We will make inference for the coefficient function based on the slope for the cumulative hazard function. We will present the coefficient functions for the transitons $0 \to 2$ and $1 \to 2$ and will compare them with the inference we reached from the previous section. The transition $0 \to 1$ will be discussed in the next section, where we will discuss methods to combine information from different transitions.

We note that the results from the previous section has been obtained after we rounded the CHD and the death times to the next year (considered them at the end of the year). The

Aalen model is fit with the original data recorded in days. When individuals have tied CHD and death times (sudden coronary death) we assumed that CHD occurred 0.5 days earlier. The Aalen model does not allow tied failures.

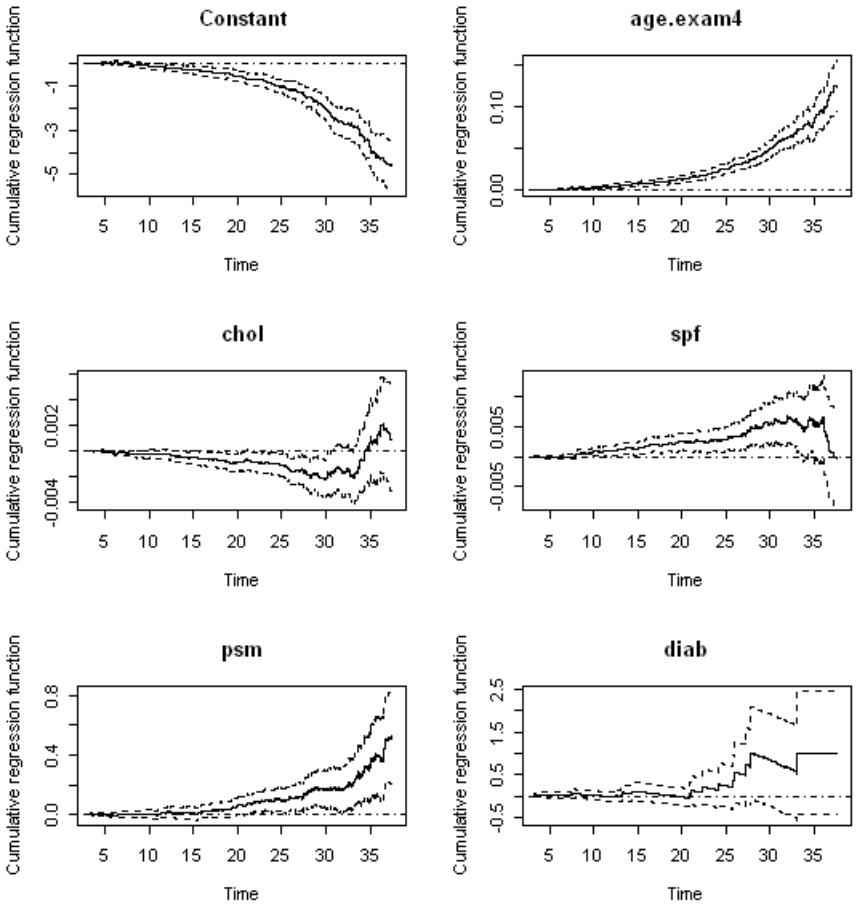

Figure 3.6: Transition $0 \rightarrow 2$, Females

Figures 3.6 and 3.7 present the 6 cumulative coefficient functions for the transitions $0 \rightarrow 2$ and $1 \rightarrow 2$ in the female group. The 95% Confidence limits for $\hat{B}_{ij}(t)$ at every time point t are included as well. We will stress again that this are confidence bounds for the cumulative function and cannot be directly used as confidence bounds for the regression functions.

The first observation is that the estimate for the cumulative regression function for *age* has an increasing slope. It increases smoothly for the transition $0 \rightarrow 2$ and resembles a piecewise linear function for the transition $1 \rightarrow 2$. This fact suggests that at the beginning

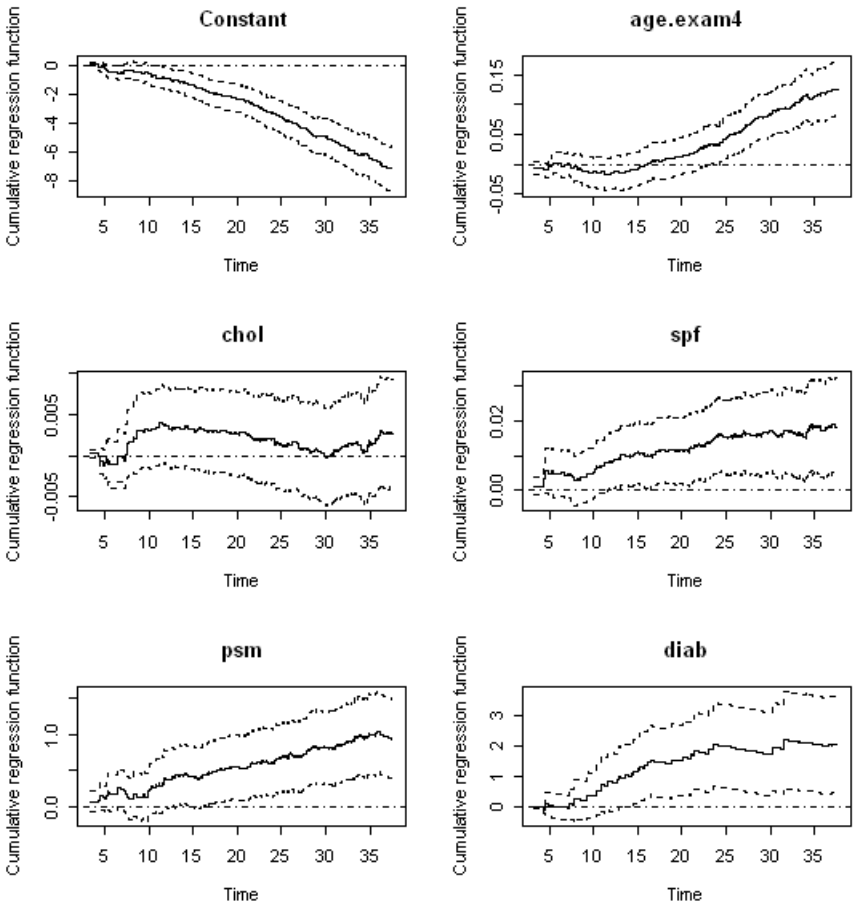

Figure 3.7: Transition $1 \rightarrow 2$, Females

of the study, the mortality of women with CHD was not affected by their age. We showed that at the beginning of the study the proportion of sudden deaths was high and the mortality rate among women with CHD was more to do with this fact, than with age. Later on, as the cohort ages, age plays more important role. We concluded in the previous section that many women with CHD appear to die of old age, i.e. the mortality is driven by the aging process and CHD plays a less important role. The piecewise linear shape of the cumulative regression coefficient for age in the female group may be used to explain why we needed the coefficient β_{age} to be a step function in the Extended Markov model. However, as it can be seen from the the 95% confidence bound, the estimate has larger variance that the estimate for the transition $0 \rightarrow 2$. The confidence bound $B_{age}(t)$ for males is narrower for

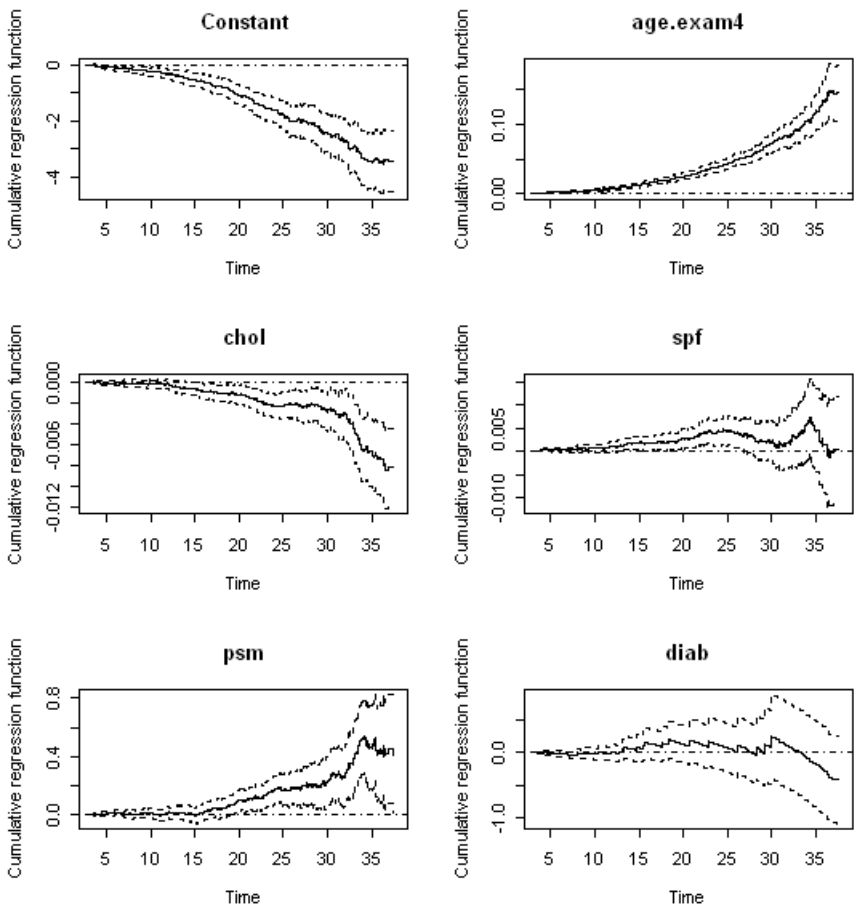

Figure 3.8: Transition $0 \rightarrow 2$, Males

the transition $1 \rightarrow 2$, perhaps due to the higher proportion of men getting the disease.

The plots of the cumulative regression functions for cholesterol are consistent with the results from the Extended Markov models in both genders. The covariate *chol* was not significant for women and appeared to be protective for males (negative coefficient β_{chol}). The covariate *diab*, which is an indicator for having diabetes at baseline was found to have stronger effect in females in the previous section. This can also be observed from the slope of the corresponding cumulative regression functions. For the transition $0 \rightarrow 2$ in females, the slope seems to be initially zero, while in the transition $1 \rightarrow 2$ is steeper at the beginning. This can be just due to the later onset of CHD in women. In both Extended Markov models, the covariate *diabX*4, which is the indicator for diabetes only for the last time period was

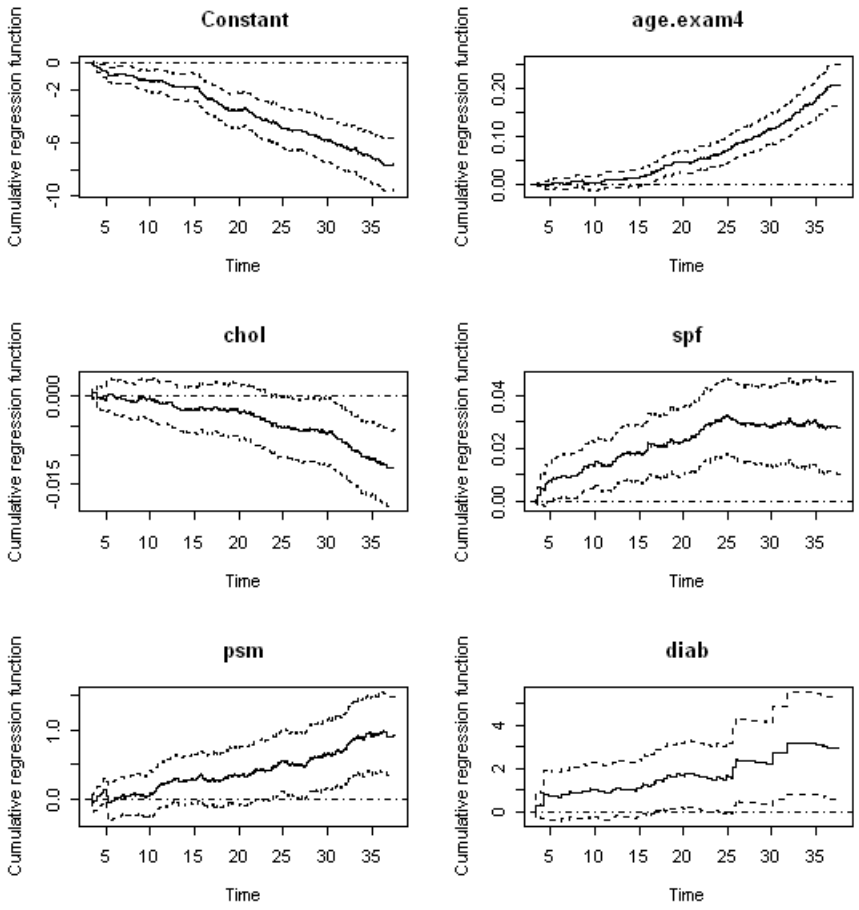

Figure 3.9: Transition 1 → 2, Males

found to be non-significant.

3.4.4 Modeling the transition to CHD

In this section we will model the transition 0 → 1. If a person dies before developing CHD, he/she is considered censored. We will use both the Cox and the Aalen models to model this hazard rate. The covariates z and c_risk, will not be used in this section, since they are always zero for this transition. Figures 3.10 and 3.11 present the results from the Aalen model. The estimated coefficients from the Cox model are given in Table 3.4.

The coefficient for age in the model for females is not significant. The reason for this finding is that the CHD is a disease that peaks at the age interval 60-70, (mean values 65.52

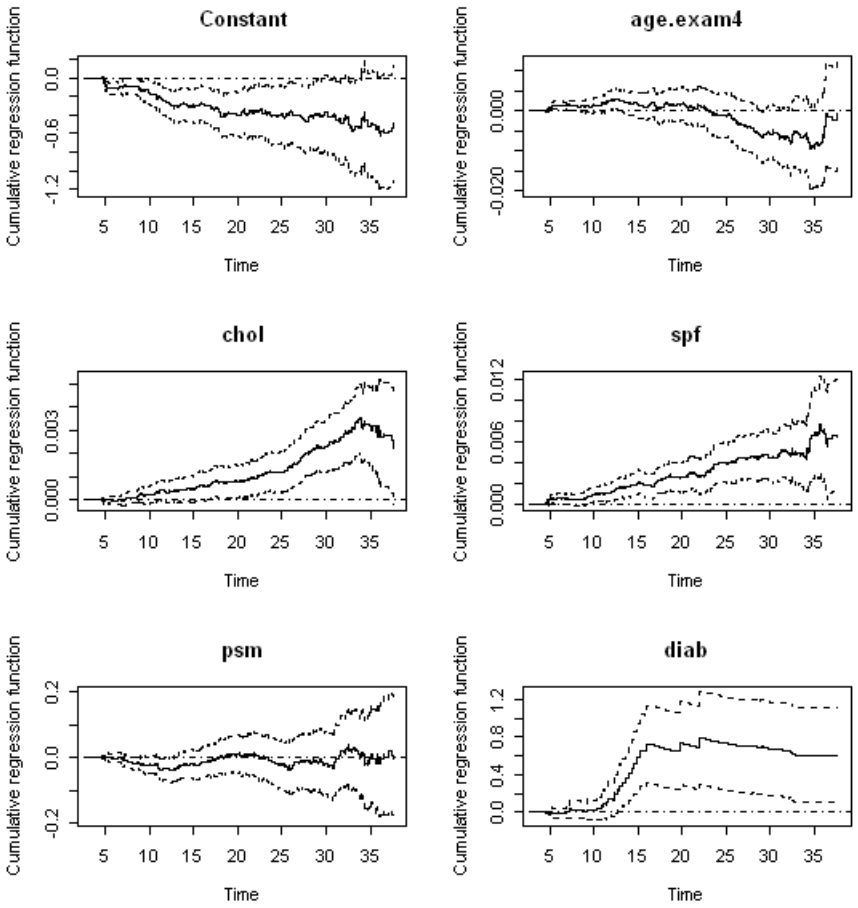

Figure 3.10: Transition $0 \rightarrow 1$, Females

for males and 69.28 for females). Initially, at the beginning of the study, the mean age of our cohort is 51.3 and older people tend to have more transitions to state 1. Later on, when the mean age of the cohort passes the age at which CHD peaks, more younger people tend to transition to state 1. The Aalen model can help us understand this phenomena. The regression coefficient function for age (as seen as the slope of the cumulative regression function) is positive initially, and somewhere around 20 years becomes negative. The time at 20 years of follow up corresponds to mean age for females of around 71.

The finding that the coefficient for age in the Cox model for women is not significant has several possible explanations. One is that other factors have more influence for whether a woman gets CHD or not. The second reason is that the Cox model has the ability of

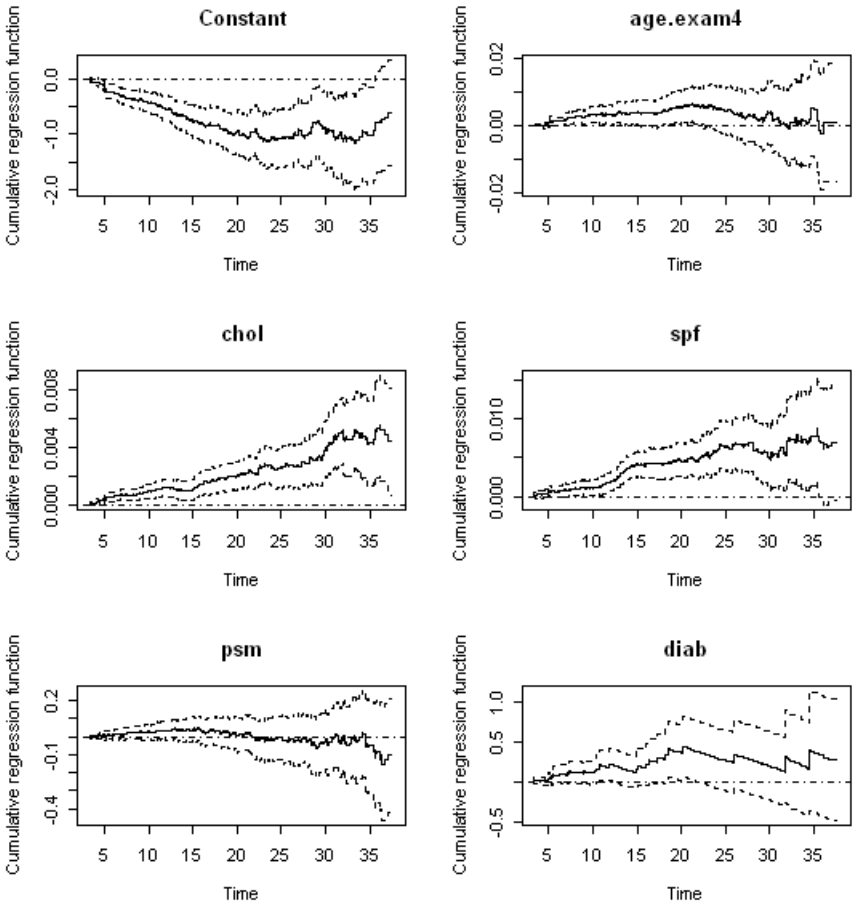

Figure 3.11: Transition 0 → 1, Males

summarizing the effect of a variable for the whole interval in the study. If at the beginning age is associated with a higher hazard of getting CHD, but later on with lowering it, the Cox model combines these effects into one coefficient. The third reason, we can think of, is that there are smaller percentage of women getting CHD than men and the test of significance has lower power. Lastly, aggressive prevention programs were organized during the time of the Framingham study. It is possible, they were more focused on a particular age group, and therefore confounded the effect of age with that of the program.

Similar results are found for the males. The coefficient for age in the model for males is significant but with a small magnitude (0.013), compared with the coefficient for the transition to state 2 (0.77) from Table 3.2. The Aalen model suggests that the coefficient

Table 3.4: 3 state model, transition to CHD state

Model for Females, transitions $0->1$						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age_exam4	.0042861	.0057243	0.75	0.454	-.0069333	.0155055
spf	.0099403	.0016677	5.96	0.000	.0066717	.0132089
diab	1.084281	.200543	5.41	0.000	.6912239	1.477338
chol	.0037301	.0009114	4.09	0.000	.0019438	.0055164
Model for Males, transitions $0->1$						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age_exam4	.0132303	.0047431	2.79	0.005	.003934	.0225266
spf	.0085506	.0018134	4.72	0.000	.0049965	.0121048
diab	.471877	.1874982	2.52	0.012	.1043873	.8393666
chol	.0044723	.0008736	5.12	0.000	.0027602	.0061845

function for age becomes negative around the same time as in the female group. We decided to keep the coefficient for age in the female group, as we are going to use it for calculating the transition probabilities in the next section.

The coefficients for cholesterol in both groups are significant, indicating 5% increase hazard (for males) for a difference of 10 mg/dL (for males). The cumulative regression function in the Aalen model has roughly a constant slope, except for the end of the study, where the risk set is small.

The coefficient for the variable *diab*, indicator for diabetes, is significant for both males and females. Diabetes increases the hazard of getting CHD by a factor of 2.94 in women, and 1.60 in men. It is interesting to look at the graph of the cumulative hazard function in the Aalen model, particularly in the female group. The slope is close to zero initially, later on rises up very sharply and again levels off. If we look in the diabetes cases, there are 45 women with diabetes, with 32 in the risk set (here risk set is alive, without CHD) at 3,578 day of follow up (9.8 years) . By the 5, 578 day of follow up (15.2 years), there are 12 left in the risk set. We can observe the ability of the Aalen model to pick up the effect of a covariate locally, but it comes as a shortcoming as well. On the other hand, the Cox model summarizes the effect of diabetes for getting CHD.

The effect of systolic blood pressure is significant for both genders, with hazards ratios 1.10 (females) and 1.09 (males) for an increase of 10 mmHg.

3.5 Transition Probabilities

3.5.1 Transition Probabilities based on the Cox model

We defined the transition probabilities in Chapter 2:

$$P_{ij}(s,\, t) = P(X(t) = j | X(s) = i)$$

where i, j are two states and $s \leq t$ are two points in time.

In this section we will focus on combining the estimated hazard rates from the individual transitions to obtain transition probabilities. This is sometimes called a synthesis of a multi-state model [28, 26].

In the simplest case, given the individual's set of covariates, we can estimate $P_{01}(0,\, t)$ and $P_{02}(0,\, t)$ - the probability that by time t, the subject will have developed CHD or have died.

In a more general situation, we may be provided with the information whether the individual has entered the CHD state by time s. The transition probabilities $P_{02}(s,\, t)$ and $P_{12}(s,\, t)$ will compare the individual prognosis, conditional on the information up to time s.

We derive the formulas for the transitional probabilities, similar to the ones derived for the competing risk model in Chapter 2.

The probability of not leaving state 0 is:

$$P_{00}(s,\, t) = exp\left(-\int_s^t (\alpha_{01}(u) + \alpha_{02}(u))d(u)\right) \tag{3.2}$$

The transition probability P_{01} in the Markovian case is:

$$P_{01}(s,\, t) = \int_s^t P_{00}(s,\, u)\alpha_{01}(u)P_{11}(u,\, t)d(u) \tag{3.3}$$

where

$$P_{11}(u,\, t) = exp\left(-\int_u^t \alpha_{12}(x)d(x)\right) \tag{3.4}$$

The transition probabilities $P_{12}(u,\, t)$ and $P_{02}(u,\, t)$ can be estimated

$$P_{12}(u,\, t) = 1 - P_{11}(u,\, t), \;\; P_{02}(u,\, t) = 1 - P_{01}(u,\, t) - P_{00}(u,\, t) \tag{3.5}$$

Another way to estimate $P_{02}(u,\, t)$ is

$$P_{02}(s,\, t) = \int_s^t P_{00}(s,\, u)\alpha_{01}(u)P_{12}(u,\, t)d(u) + \int_s^t P_{00}(s,\, u)\alpha_{02}(u)du \tag{3.6}$$

In the above formula, the first term corresponds to probability of going to state 2 via state 1, and the second term, to the probability of going directly to state 2. The two expressions for $P_{02}(s,\, t)$ are equivalent.

If we have the extended Markov model then $P_{11}(u,\, t)$ above and in Equation 3.3 has to be replaced by

$$P_{11}(u,\, t) = exp\left(-\int_u^t \alpha_{12}(x,\, x-u)dx\right) \tag{3.7}$$

In section 3.4 we developed models using the variables z- an indicator for being in the CHD state and c_risk- an indicator for just entering the state, and d- time spent in state 1. Further, we interacted these variables with the follow up time, using 4 periods. Therefore, we modeled the effect of CHD as changing the hazard of dying by a factor, depending on when the transition occurs and the the time spent with the disease. In this section, in order to estimate the transition probabilities, we need to estimate the baseline hazard, which presents problems when using time-varying covariates both computational and inferential, as discussed in [26] (p.307).

To avoid this, we can adopt a different approach. This will present another use of the multi-state models. Instead of assuming that the hazard changes by a factor, we can assume that the effect of some factors is altered moving to another state, e.g. the effect of high blood pressure is different before and after getting CHD. This corresponds to assuming different coefficients for covariates before and after getting CHD. This approach is used by Klein at. al. [26] for modeling the survival of cancer patients after bone marrow transplant.

We will use the following convention: for a variable v, we will denote $v\_nochd = vI[t \leq u]$ and $v\_postchd = vI[t > u]$, where u is the time of the diagnoses of CHD, and $u = \infty$ for individuals who never visit the CHD state.

The transition $0 \rightarrow 1$ was modeled in section 3.4.4. Next, we will present the results for the transitions $0 \rightarrow 2$ and $1 \rightarrow 2$ based on the approach outlined above. The results are presented in Table 3.5.

In both models, only the effect of diabetes after CHD is found to be significant. In both models, the effect of systolic blood pressure (variable spf) is different for people with CHD.

Table 3.5: 3 state model, Females

Model for Females, transitions $1->2$ *and* $0->2$						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age_exam4	.0699832	.0039429	17.75	0.000	.0622553	.077711
psm	.3391367	.0577513	5.87	0.000	.2259463	.4523272
spf_nochd	.0051733	.0011716	4.42	0.000	.002877	.0074695
spf_postchd	.0085612	.0011441	7.48	0.000	.0063188	.0108035
diab_postchd	1.31061	.204432	6.41	0.000	.9099308	1.71129
Model for Males, transitions $1->2$ *and* $0->2$						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age_exam4	.0769317	.003771	20.40	0.000	.0695407	.0843228
psm	.3844864	.0658804	5.84	0.000	.2553632	.5136096
spf_nochd	.0059262	.0014694	4.03	0.000	.0030462	.0088061
spf_postchd	.0129554	.0014401	9.00	0.000	.0101328	.015778
diab_postchd	.9668435	.1929836	5.01	0.000	.5886026	1.345084
chol	-.0027605	.0007303	-3.78	0.000	-.0041919	-.0013291

We want to address two issues before moving further. We have already discussed the problem of using the covariate z for calculating the transition probabilities. However, the covariate z was not found to be significant, after introducing the covariates in the Table 3.5. In a way, that could be explained as the effect of z is now explained by the different effects of some of the factors before and after the disease.

The second question that arises is whether the models in section 3.4.2 could have benefited from introducing interaction terms with the covariate z. If the interaction term $zXdiab = zdiab$ is included in the models from Table 3.3, both *diab* and $zXdiab$ are not significant, so we have to choose one of them. We decided to use *diab*, since the number of people with diabetes is very small.

The variable $zXspf$ is not significant in males.

Next, the models from Tables 3.5 and 3.4 are used to calculate the transition probability $P_{02}(s, t)$ using equation 3.6.

3.5.2 Plots of the Transition Probabilities

Transition Probabilities can be used to estimate the prognosis of dying for an individual with a particular set of covariates. In order to calculate them, we need to fix the covariates.

For the continuous covariates, we considered the values $age\_exam4 = 51$, $chol = 241$ and $spf = 150$. The categorical variables have different values, as indicated on the plots. The transition probabilities can also be used to compare the probability of dying, conditional on whether they have developed CHD or not by a certain time. We give two examples with, estimating $P_{02}(s, t)$ for $s = 11$ and $s = 21$ years. Figures 3.12 and 3.13 present the transition probabilities $P_{02}(s, t)$ for $s = 11$ and $s = 21$ years for a non-diabetic woman. The two graphs are for a patient with and without CHD at the time s of the prediction. Similarly, figures 3.14 and 3.15 illustrate the same graphs for a diabetic woman.

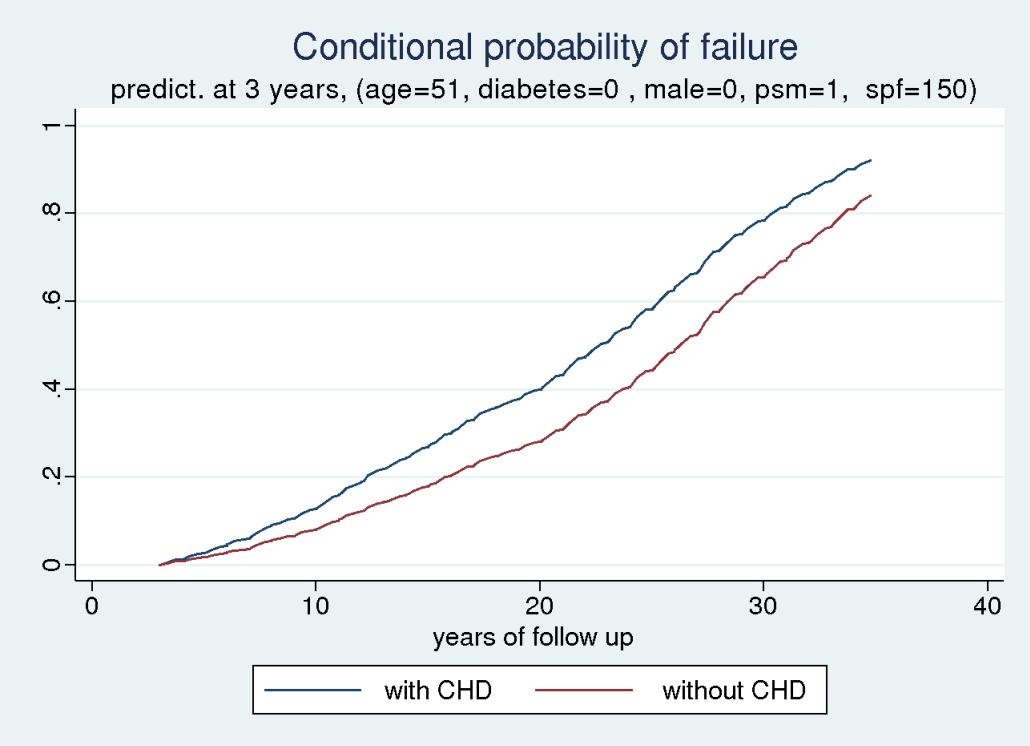

Figure 3.12: Transition Prob. to state 2, Females, non-diabetic

3.5.3 Transition Probabilities based on the Aalen model

Estimated hazard rates using the Aalen model can be easily combined to obtain transition probabilities. For a given covariate vector $\boldsymbol{X}^0$, the cumulative hazard rates are estimated

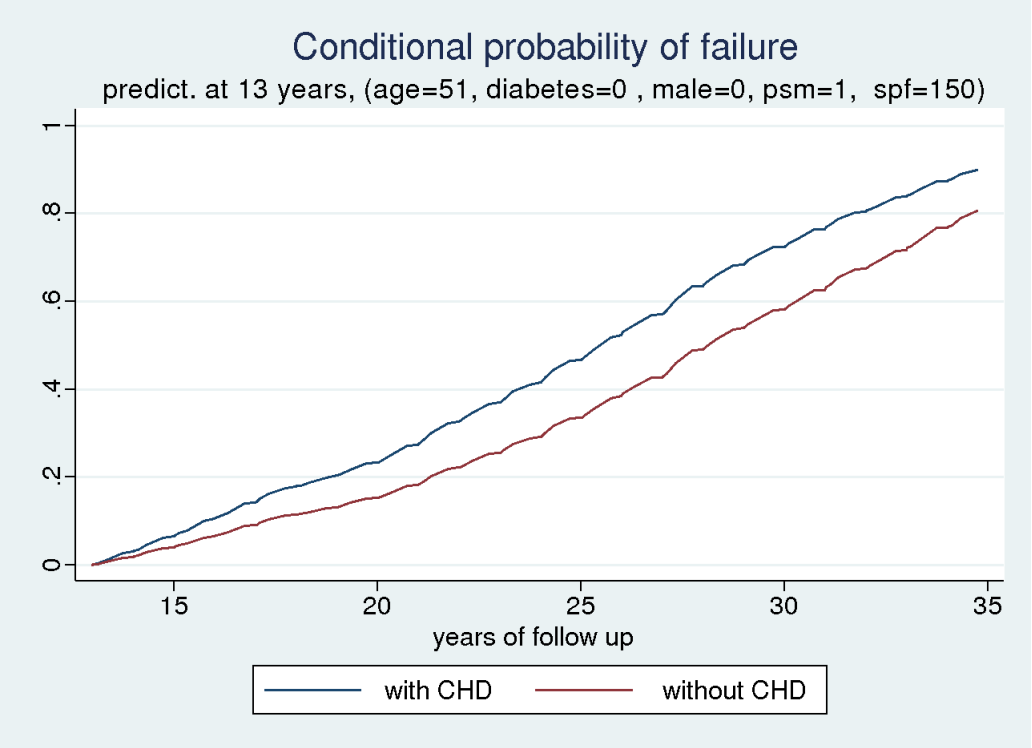

Figure 3.13: Transition Prob. to state 2, Females, non-diabetic

by:

$$\hat{A}_{ij}(t, \boldsymbol{X}^0) = \int_0^t \boldsymbol{X}^0_{ij}(s)' d\hat{\boldsymbol{B}}_{ij}(s)$$

As in the case of Aalen-Johansen estimator, presented in Chapter 2, the product integral can be used to obtain estimates for the transition probabilities:

$$\hat{\boldsymbol{P}}(s, t) = \prod_s^t (\boldsymbol{I} + \hat{\boldsymbol{A}}(du)) \tag{3.8}$$

Figures 3.16 and 3.17 present plots of the transition probabilities $P_{02}(0, t)$ for a non-diabetic and diabetic woman. We observe similar curves as in figures 3.12 and 3.14. The transition rate $P_{02}(0, t)$, as in formula 3.6, combine direct transitions to state 2, as well as through state 1. The program Addreg was used to calculate the transition rates. The program does not allow calculating the transitional probability, conditioning on the history at time $s < t$. The plots estimate the individual prognosis for a female smoker, 51 years old at baseline with total cholesterol level of 241 and systolic blood pressure of 150.

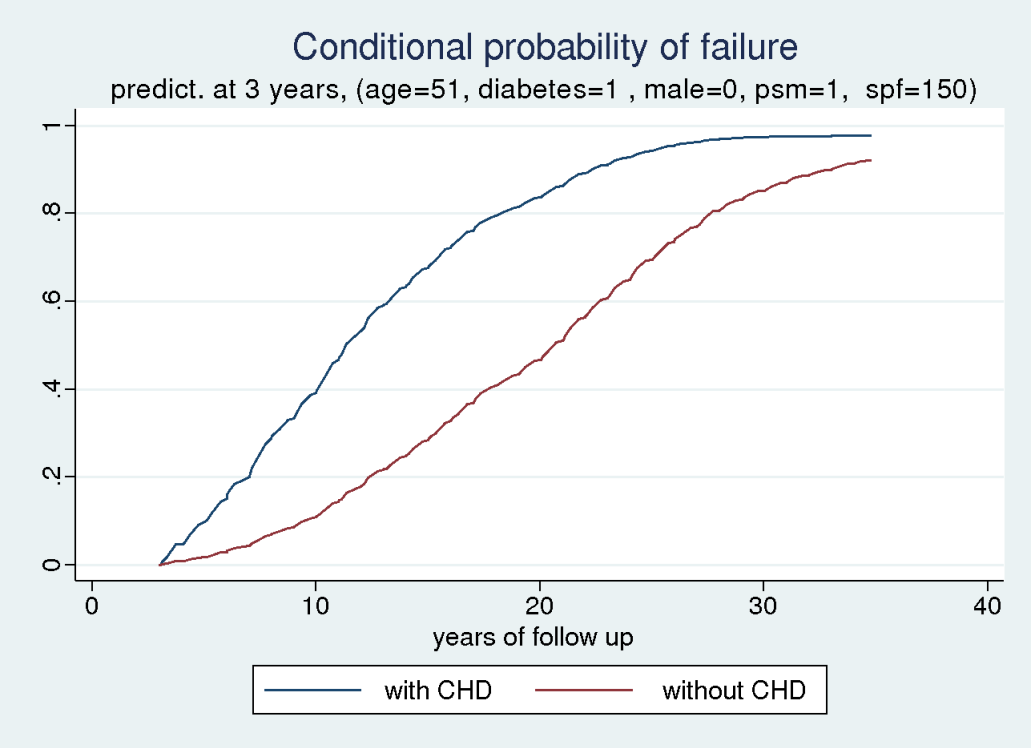

Figure 3.14: Transition Prob. to state 2, Females, diabetic

The plots in figures 3.18 and 3.19 present the transition probabilities of moving to the CHD state for a diabetic and non-diabetic woman. The other covariates are set as described above. As we found earlier, initially this probability increases, and later on as the subjects age, the probability decreases. The increase at the beginning is steeper for females with diabetes. The models we developed in section 3.4.4 revealed that diabetes highly increased the risk of CHD, especially for women.

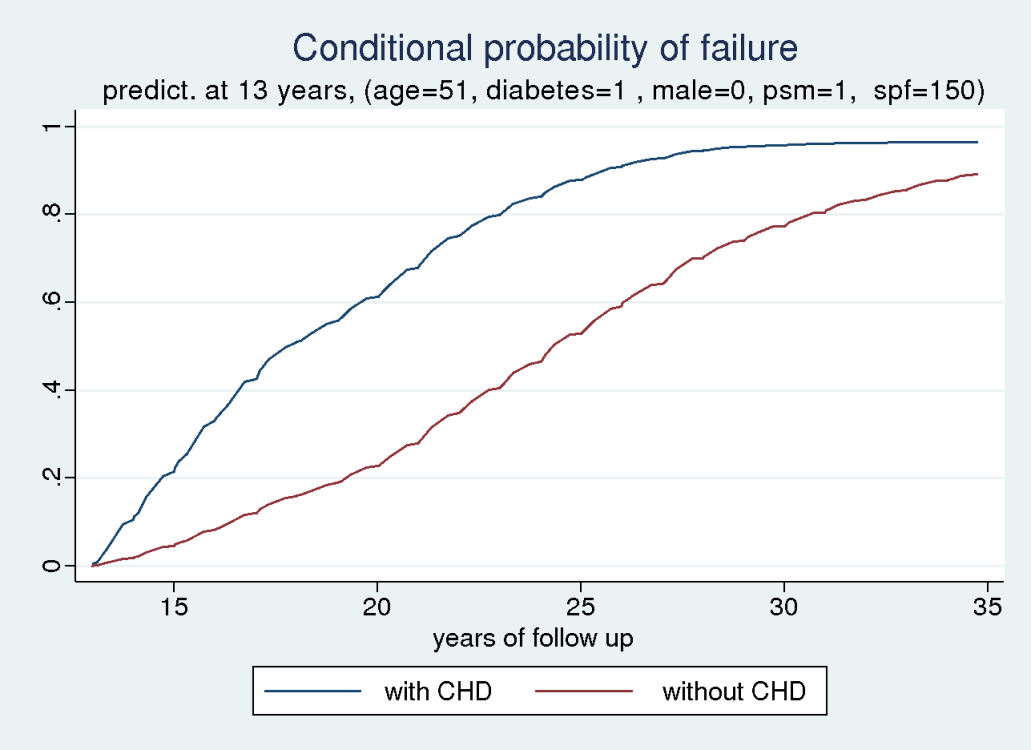

Figure 3.15: Transition Prob. to state 2, Females, diabetic

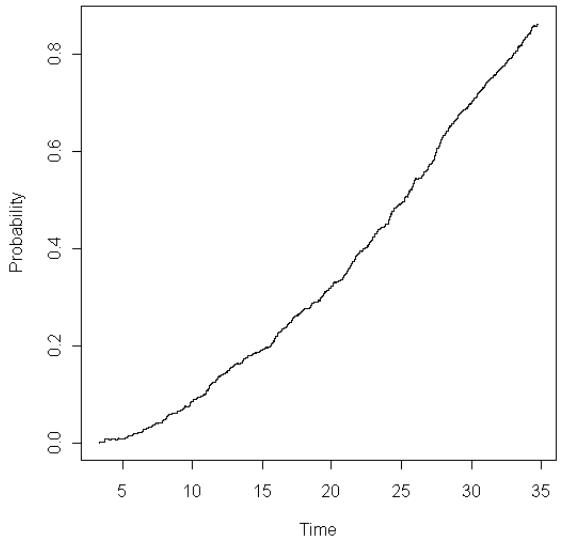

Figure 3.16: Transition Probabilities,$P_{02}(0,t)$, Female, non-diabetic

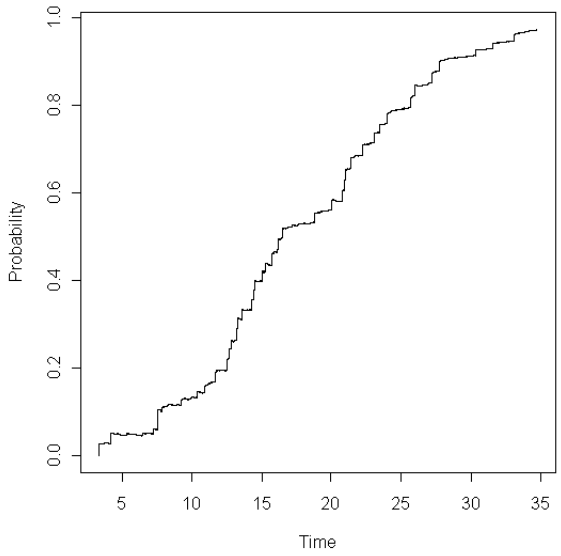

Figure 3.17: Transition Probabilities, $P_{02}(0,t)$, Female, diabetic

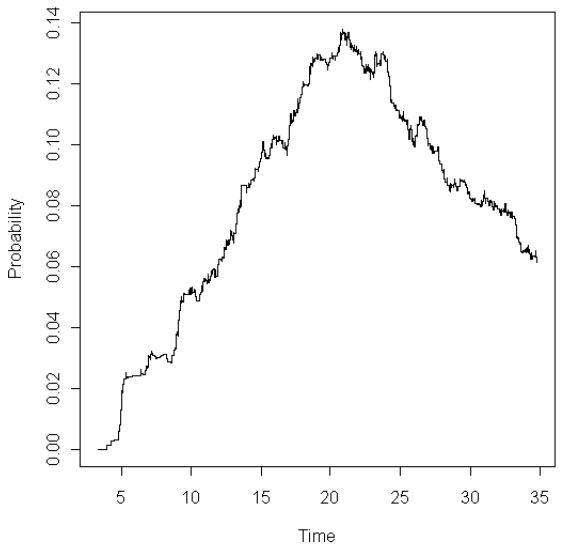

Figure 3.18: Transition Probabilities, $P_{01}(0,t)$, Female, non-diabetic

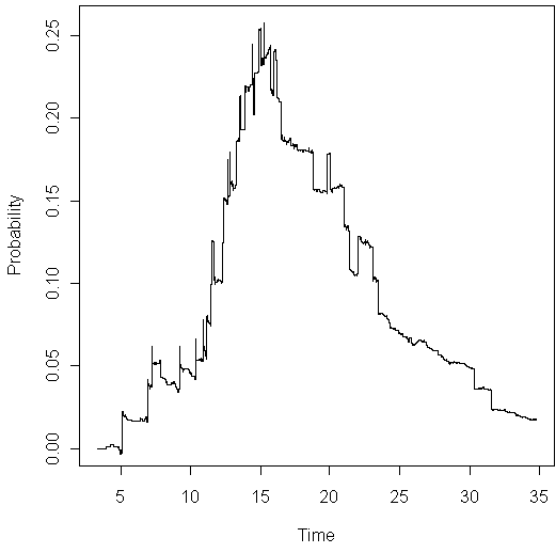

Figure 3.19: Transition Probabilities, $P_{01}(0,t)$, Female, diabetic

CHAPTER 4

Five-state Models

It is estimated that by year 2020 heart disease and stroke will become the leading cause of death and disability world wide. In this chapter we will study the effect of CHD (Coronary Heart Disease) and CVA (Cerebral Vascular Accident) individually and jointly on subsequent mortality. We will be interested in whether the occurence of either disease makes it more likely to develop the other, and whether the role of the covariates changes depending on the current state. This is a more realistic examination of the role of chronic diseases in an aging population and involves considering 5 states.

We use a multi-state model presented in Figure 4.1. This model has the following five states:

- State 0 - The "healthy" state. All participants begin in this state, free of both CHD and CVA.
- State 1 - The CHD state. Participant developes CHD prior to developing CVA.
- State 2 - The CVA state. Participant developes CVA prior to developing CHD.
- State 3 - The state signifying that a participant has developed both CHD and CVA.
- State 4 - The Death state.

We use the data set from the Framingham Heart Study that was introduced in Chapter 3. As before, the dataset contains information 3,201 individuals (1,681 females) with 2,722 deaths (1,402 in females). The five state model we consider allows 8 possible transitions which we denote: $01, 02, 04, 13, 14, 23, 24, 34$. I.e., the transition ij occurs when a subject moves from state i to state j.

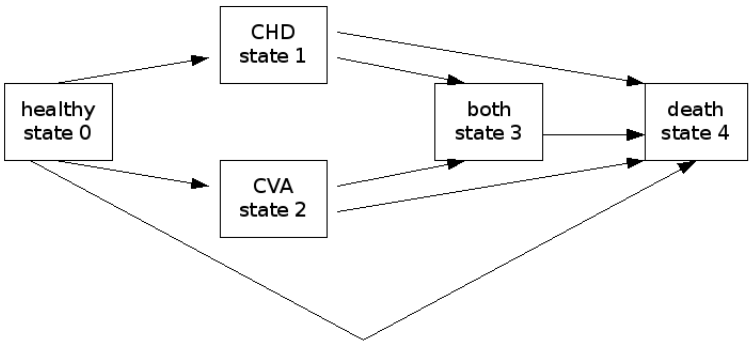

Figure 4.1: Five state model

Table 4.1: Direct transition to each state

state	No of visits	Direct transitions to each state				
		0	1	2	3	4
0	3,201	83	1,307	509	0	1,302
1	1,307	0	240	0	191	876
2	509	0	0	101	93	315
3	284	0	0	0	55	229
4	2,722	0	0	0	0	2,722

Table (4.1) presents the number of observed direct transitions. At the beginning of the study, all subject were in state 0. In the table, the numbers representing the transition ii are the number of the subjects who, once in that state, remained in the state. For example, there were 55 subjects who were observed to have the transition 33. They are considered censored in state 3 (had both CHD and CVA, but did not die by the end of the study). There were 83 subjects (3%), for whom no transition from state 0 was observed.

The total number of people suffering from CVA is the sum of the counts for the direct transitions $0 \rightarrow 2$ (509) and $1 \rightarrow 3$ (191), total of 700 (391 for women).

State 3 is visited only by individuals who experience both events-CHD and CVA. However

subjects can move to state 3 from either state 1 or 2, depending which event occurs first. For example, a disease progression for an individual who suffers first from a stroke, later develops CHD and dies before the end of the study will be described by the multi-state model with the transitions 02, 23 and 34. From the 1,307 individuals who visited state 1 (developed CHD), 191 (15%) moved to state 3 (developed CVA). Similarly, from the 509 visits to state 2 (developed CVA), 93 (18%) later transitioned to state 3 (developed CHD).

4.1 Cerebrovascular accidents (CVA) and CHD

Cerebrovascular accident (CVA), commonly known as a stroke, is the death of brain tissue due to the loss of blood flow to a particular area of the brain. Blood flow may be lost due to a blockage or the rupture of a vessel. First, we want to compare CVA with CHD, later we will examine how these two diseases affect mortality.

The mean age of CHD diagnoses are 65.5 years among men and 69.2 years among women. Similarly, the mean age of CVA diagnoses are 70.1 years for men and 74.1 years for women. In both diseases, the disease occurs, on average, about 4 years earlier among men than among women. And, the average age of diagnosis for CVA happens later in life that the average age of diagnosis of CHD in both genders. This can be seen in Table 4.1 as well, since from the 284 subjects visited state 3, 191 (67%) transitioned from state 1, i.e. first experienced CHD.

When we examined the occurrence of CHD alone in Chapter 3, we noted that a significant number of the deaths after having CHD occurred within a day of the diagnosis and termed these "sudden coronary deaths." In Chapter 3 we also observed that the proportion of sudden deaths was decreasing over the time-span of the Framingham study. A major difference between the two diseases is in the distribution for the time between the occurrence of the disease and death. Overall, from the 1,105 subjects who died after developing CHD, sudden deaths account for 20% of the deaths . On the other hand, from the 544 individuals who died after developing CVA, only 2% died within a day of the diagnosis. The difference between the distributions may be seen by comparing the three quartiles (25%, 50%, and 75%) for the time between the occurrence of disease and mortality for the two diseases. For CHD these quartiles are 19, 1,761, and 4,453 days; while for CVA the quartiles are 67, 1,110, and 3,011 days.

We discussed models for predicting the occurrence of CHD in Chapter 3. Here, we first focus on building prognostic models for CVA. We will then examine whether developing one disease (CVA or CHD) makes it more likely for a subject to develop the other.

The models for developing CVA for females and for males are presented in Table 4.2. To arrive at this model, we began with a candidate set of variables, age, cholesterol, smoking status, systolic blood pressure, and diabetes. We used likelihood ratio statistics to determine which variables made a significant contribution to the model.

Table 4.2: Model for developing CVA (transition 02).

Model for Females, transitions to CVA						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf. Interval*	
age	.0377681	.0071433	5.29	0.000	.0237674	.0517688
spf	.0130581	.0020989	6.22	0.000	.0089444	.0171718
Model for Males, transitions to CVA						
age	.0414167	.0076317	5.43	0.000	.0264588	.0563746
psm	.2904497	.1351825	2.15	0.032	.0254969	.5554024
spf	.0162012	.0028081	5.77	0.000	.0106973	.0217051
diab	.8578151	.2674009	3.21	0.001	.333719	1.381911

Different sets of covariates affect the rate men and women develop CVA. The only covariates found to increase this rate in women are age and blood pressure. The model for males includes diabetes as well as the smoking status. Both the effects of age and blood pressure are found to be stronger in men.

It would be interesting to compare the models for developing CVA those for CHD, presented in Chapter 3. The effects of age, and blood pressure are stronger in the CVA models for both genders. Cholesterol was found to be significant predictor for developing CHD, while it does not appear to play a role for CVA. Smoking is included in the CVA model for males, while it was not included in neither model or CHD. Diabetes was very strongly associated with the risk of CHD in women, and is not found as a significant covariate for developing CVA in females. On the other hand, it seems to play more important role for CVA (coefficient $\beta_{diab} = .86$) than for CHD (coefficient $\beta_{diab} = .47$) in the male group.

One question we are interested in, is whether having one disease affects the probability that individuals will get the other disease. To examine this, we compare the hazard rates for

CVA, before and after subjects have CHD. The model for CVA combines the transitions 02 and 13, i.e. having a stroke before CHD is diagnosed and after that. Let *z_chd* be a variable which is an indicator for having CHD. In order to compare the rates males develop CVA before and after CHD, we can include *z_chd* in our model. This is similar to the approach we took in Chapter 3 and corresponds to the assumption that the hazard of getting CVA increases by a factor after CHD is diagnosed. The covariate *z_chd* was not found to be significant.

However, another way CHD may influence the transition to state 3 is by altering the effect of some covariates. To investigate this, we add the interaction of the covariates from Table 4.2 with *z_chd*. None of these variables were found to be significant. Therefore, we concluded that the rate at which males are having strokes is the same before and after CHD is diagnosed. Similarly, we compared the rates individuals are developing CHD, and having CVA was not found to have any effect. The last finding is less important, since as we discussed earlier, people usually develop CVA later in life.

4.2 Mortality models using the five-state model.

Next, we will use the 5 state model we described earlier to investigate the role of CVA and CHD in subsequent mortality. There are four transitions to state 4:

- 04, a person may die without having developed CHD.
- 14, a person may die after having CHD, but without having developed CVA.
- 24, similarly a person may die after having CVA without having developed CHD.
- 34, a person may die after having developed both CHD and CVA.

The goal is to compare the hazard rate for these different transitions and to determine which risk factors affect these hazard rates. As in Chapter 3, we will consider separate models for males and females. First, we will compare different models for males, later we will present a model for females.

4.2.1 Mortality models for men, Markov and non-Markov models

The simplest model we may assume is the Markov model that assumes that the probability of transition at any time is a function only of time and of the present state. To model this

assumption we can indicate the effect of being in states 1, 2 and 3 by including indicator variables for being in these states in our model. The indicator variables are denoted z_chd, z_cva and z_both for being in states the CHD only, the CVA only, and the CHD and CVA state respectively.

We examine two possible departures from the Markov assumption. It is possible that the hazard rate depends on the past history of the process or that the hazard depends on the time spend in a particular state .

For example, we noted in Chapter 3, that mortality is very high after moving to the CHD state. In our current model, there are two transitions possible after developing CHD, to state 3 if the person developes CVA or directly to state 4 if the person dies without developing CVA. We compared the hazard rate from state 3 for the individuals, who were diagnosed with CHD when moving to the state 3, i.e. those moving through the transitions 02, 23.

It may be conjectured that the hazard rate from state 3 depends on the time spent with any disease (time spent in state 2 or 3). For example, it may be expected that developing both diseases in a short interval of time is associated with higher mortality, compared to those with a longer time between the diagnoses. There are other possible ways the hazard may depend on the event history. For example, the hazard rate from state 3 to state 4 may depend on the time spent with CHD or CVA or on which disease developed first.

Examining these possibilities, we found that the hazard rate from state 3 is higher for individuals diagnosed with CVA first, but does not depend on the time spent in either state 1 or 2. In our analyses, we introduced the new covariate $cvafirst$, an indicator for having CVA before CHD and found it was a significant determinant of the hazard for moving from state 3 to state 4. This covariate affects only the transition $3 \rightarrow 4$ and makes the model non-Markov, since the hazard depends on previous states as well as the current state. The result is presented in Table 4.3.

The coefficient for cholesterol is very close to what we obtained for the models in Chapter 3. The same is true for the coefficients for age, systolic blood pressure, smoking status and diabetes. The hazard of dying increases by a factor of 2.89 if only CHD is present and by a factor of 2.27 if only CVA is diagnosed. If both are present, the hazard of dying depends on which disease was diagnosed first. If CHD is first (more common), the hazard is increased by a factor of $\exp(1.06 + .82 - .42) = 4.31$. If CVA is first, the hazards increases by $\exp(1.06 + .82 - .42 + .45) = 6.75$. In both cases, state 3 has a mortality rate higher than

Table 4.3: Mortality model , Males

Mortality model for Males						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\lvert z \rvert$	*95 % Conf.*	*Interval*
age	.0782033	.0038345	20.39	0.000	.0706877	.0857189
psm	.3655678	.0657504	5.56	0.000	.2366994	.4944362
sbp	.0075643	.0014278	5.30	0.000	.0047659	.0103628
diab	.4199214	.1472968	2.85	0.004	.131225	.7086179
chol	-.0027346	.0007382	-3.70	0.000	-.0041815	-.0012878
z_chd	1.061675	.0656575	16.17	0.000	.9329886	1.190361
z_cva	.8215671	.1049697	7.83	0.000	.6158303	1.027304
z_both	-.4229321	.1604893	-2.64	0.008	-.7374854	-.1083788
cvafirst	.4588426	.1824337	2.52	0.012	.1012792	.816406

from the other states. This can be observed in Table (4.1) - from the 284 subjects (both genders) visiting state 3, 229 (81%) failed.

The test of the proportionality assumption reveals that it is not satisfied for the indicator variables for states 1, 2 and 3. In order to avoid this problem we can interact these variables with time intervals, as we did in Chapter 3. Instead we will look for a different approach, which we present in the models that follow.

4.2.2 Models for Males with different baseline hazard rates.

In order to develop a model which satisfies the proportionality assumption, we will allow different hazard rates to have different baseline hazards. The effects of the covariates will also be allowed to depend on the transition. This corresponds to modeling the hazard as follows:

$$\alpha_{i4}(t|\boldsymbol{X}) = \alpha^0_{i4}(t)\exp(\boldsymbol{X}_{i4}\boldsymbol{\beta}_{i4})$$

where $i = 0, 1, 2, 3$ and $\alpha^0_{i4}(t)$ is the baseline hazard for the transition $i4$.

This corresponds to a stratified Cox model, where the mortality rate from each state is in a different strata. For example, the hazard rate for the transition 14 corresponds to strata 14.

In this approach, we cannot use indicator variables such as *z_chd*, since they are constant for each strata. The increased mortality for moving to state 1 for example, would reflect in

a higher basaline hazard rate from this state. A model using this approach is presented in Table 4.4.

Table 4.4: Mortality model , Males

Model for Males, transitions to CVA						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\|z\|$	*95 % Conf. Interval*	
age	.0835918	.0049735	16.81	0.000	.0738439	.0933396
*age*14	-.0243457	.007741	-3.15	0.002	-.0395177	-.0091737
psm	.3589073	.0664419	5.40	0.000	.2286836	.489131
spf	.0069044	.0014519	4.76	0.000	.0040586	.0097501
*diab*14	.9352022	.2172404	4.30	0.000	.5094189	1.360986
chol	-.0028859	.000741	-3.89	0.000	-.0043382	-.0014336
$cvafirst$	.3502635	.1879008	1.86	0.062	-.0180153	.7185423

The coefficients for smoking, cholesterol and blood pressure are found to be very close to the ones from the model in Table 4.3. The effect of age reduces for the hazard rate 14. The estimated coefficient for this hazard is $\beta_{14}(age) = .084 - .024 = .060$. One explanation for the reduction in the effect is due to the high rate of mortality after CHD in men. The high proportion of deaths immediately after CHD increases mortality for the strata 14 and reduces the role of age. In other words, for two men , given they both have developed CHD, the hazard ratio for one year difference in age is $exp(.06) = 1.06$ or 6% higher. The hazard ratio before they developed CHD (or generally in the same strata, different from 14) is $exp(.084) = 1.09$ or 9% higher.

The high proportion of deaths immediately after CHD, we believe, makes the effect of the covariate $cvafirst$ almost significant (p-value from the Wald test is 0.06). Men who enter state 3, having experienced CVA before, are those having CHD at the time. Their mortality is higher compared to those, who developed CHD first, survived the initial hight mortality rate and later developed CVA, at which time they enter state 3.

Another possible contribution to the covariate $cvafirst$ is the later onset, on average, of CVA compared to CHD. In other words, as we discussed at the beginning of Chapter 4, CVA occurs on average 4 years later than CHD. Similarly, the average age for males, entering state 3 from state 2 is 73.5 years old (55 observations), versus average age of 71.5 years old (95 observations) for those having CHD first. We will examine the contribution of this difference

later.

The coefficient for diabetes is significant only for the hazard rate 14. It is larger that the one in Table 4.3.

The proportionality assumptions are satisfied for all variables, except for spf and $age14$. This is a finding we also observed in Chapter 3. The effect of blood pressure is reduced later in the study and we can account for this allowing a different coefficient for the last 15 years of the study, as we did in Chapter 3. There may be several reasons for this phenomena. Blood pressure is measured at baseline, and in general blood pressure is known to vary more than other covariates. Another reason is the use of effective medication for hypertension made available in the second half of the study's follow up. A third reason may be that the effect generally reduces with age, however in this particular study, this effect is confounded with the effect mentioned previously.

The effect of $age14$ also reduces with time, and we believe has to do with the reduced rate of mortality immediately after CHD. The higher the mortality rates are, the less important the effect of age is, and therefore the larger in magnitude the coefficient $\beta_{14}(age)$ would be.

In Chapter 3, we considered coefficients for age and blood pressure to be piecewise constants, allowing the effect to change with time. We adopted this approach, since our goal was to compare hazard rates before and after CHD. The same approach can be applied here, but would be more tedious, since we have 8 hazard rates instead of 3. We will use a different approach which will be illustrated in the next section.

We want to look into the effect of the variable $cvafirst$ more closely. Earlier we hypothesized that some of the effect of the covariate may have to do with the fact that on average CVA is developed 4 years later than CHD (2 years for males entering state 3). We will consider a model only for the hazard rate $\alpha_{34}(t|\boldsymbol{X})$, i.e., only for the transition 34. This resembles the approach we will use in the next section. There we will use models, where time is measured since entering the state. For this hazard rate, the reason we consider time since entering the state, is the higher rate of mortality from state 3. From the 284 subjects who visited state 3, 229 died (81% in both genders, 84% for males). Therefore the combination of both diseases appears to present a serious health burden, and the effect of other covariates may be reduced and the time since entering the study might be most important.

The covariate age is not used, rather age_state3 is used, which is the age at entering state

3. As we discussed we want to detect if the effect of the covariate $cvafirst$ has to do with the later onset of CVA. The results are presented in Table (4.5).

Mortality models for Males, strata 34, age at entering state 3

Table 4.5: Mortality model from state 3 , Males

Model for Males, transitions to CVA						
$\_t$	$\hat{\beta}$	*Std. Err.*	z	$P>\|z\|$	*95 % Conf.*	*Interval*
age_state3	.0559122	.0123956	4.51	0.000	.0316173	.0802071
spf	.014985	.004307	3.48	0.001	.0065433	.0234266
$cvafirst$	.3467697	.1859612	1.86	0.062	-.0177075	.711247

We obtain practically the same coefficient for $cvafirst$ as previously. Only 3 covariates are found to be significant. The effect of age (age at entering state 3) is reduced. The coefficient for blood pressure is higher than the one in Table 4.4. This suggests, that possibly a separate coefficient $\beta_{34}(spf)$ could be used in the model from Table 4.4. The effect of the covariate $spf34$ was not found to be significant. The model considered above uses only a small portion of the data set, 150 observations with 126 failures. The test of the proportionality assumption indicates that it is satisfied (p-value of .75).

4.2.3 Mortality models for Males,time since entering the current state

We will consider models, where time is measured since entering the state. The importance of what constitutes time in the Cox model is related to the proportionality assumption. As we discussed in Chapter 2, the Cox model assumes that the hazard rates for two subjects are proportional, with a ratio depending on the covariates, but not on time. The hazards are compared at the same time point. If we use time since entering the study, we compare subjects at the same time after they entered the study. This has advantages, as we discussed in Chapter 3, since the coefficients are adjusted for any cohort effects. Also in Chapter 3, we briefly discussed that age has been used as a time scale [9]. If this approach is used, we assume the hazards are proportional for two subjects at the same age, when the time since they entered the study may be different.

In the previous section, we discussed a model only for the hazard rate $\alpha_{34}(t)$ and argued that due to the severity of having both CVA and CHD this approach is justified. The rationale for using time since "in state" is also advantageous for states 1 and 2 since there are increased mortality initially, especially in state 1. For the transition $0 \rightarrow 4$ time since entering the state or the study are equivalent.

A model using time since "in state" is presented in Table 4.6.

Table 4.6: Mortality model , Males

Model for Males, transitions to CVA						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>* $\lvert z \rvert$	*95 % Conf. Interval*	
age	.0772989	.0047785	16.18	0.000	.0679333	.0866646
*age*14	-.0359369	.007256	-4.95	0.000	-.0501584	-.0217155
psm	.3001082	.065809	4.56	0.000	.1711249	.4290915
spf	.0048826	.0013979	3.49	0.000	.0021428	.0076225
*diab*14	.5978204	.2163211	2.76	0.006	.1738389	1.021802
chol	-.002613	.0007358	-3.55	0.000	-.0040551	-.0011708
cvafirst	.3783978	.1840152	2.06	0.040	.0177346	.7390609

The coefficients for age are similar with the ones in Table 4.4. The coefficient for age is slightly smaller, with a larger coefficient for *age*14. The interpretation of the coefficients here is different, however. When the time was considered since entering the study, the coefficient can be interpreted as hazard ratio for one year difference in age. Here, we compare subjects at equal amount of time after they entered state 1. Therefore the hazard ratio $exp(.77 - .35) = 1.04$ suggests a 4% increase for a year difference at baseline (and not difference of age). The coefficients for blood pressure, for entering the CVA state first and for smoking are slightly reduced. The coefficient for diabetes is reduced by 57%. The test of the proportionality assumption suggests it is satisfied for this model.

4.2.4 Mortality models for Females, semi-Markov model

In the previous sections, we compared different models for males and explained how we build the final model, presented in Table 4.6. Similarly for women, we followed the same approach and compared models with indicator variables for each state and stratified models with time since entering the study as the time scale. We also found that the proportionality assumption

was violated at least for one variable. The model we have chosen as our final model, as in the case for males, uses time since entering the study. However, the role of the covariates is different.

We will present two models for women, one only for the transition 34 and a final model for all transitions. The results for a model only for the hazard 34 are presented in Table 4.7. Unlike the model for men, the covariate $cvafirst$ is not significant and cholesterol and smoking are found to be significant. The model is based on 134 observations (103) failures, time is measured since entering the study and age is measured at baseline.

Table 4.7: Mortality model from state 3 , Females

Model for Males, transitions to CVA						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age	.0572218	.0136396	4.20	0.000	.0304887	.0839549
diab	1.388561	.5292738	2.62	0.009	.3512039	2.425919
psm	.7663314	.2467098	3.11	0.002	.2827892	1.249874

The results assuming different baseline hazards for each transition are presented in Table 4.8. These models use time "in state." as the time scale. As in the model for the transition 34, the model does not require the covariate $cvafirst$. As we discussed earlier, we believe that this covariate captures the effect of high mortality immediately after CHD, which is more present for males, as we found in Chapter 3.

Covariates measuring time spent in states 2 or 3 does not affect the mortality rate from state 3. Therefore, this is a Semi-Markov model, since the hazard depends on the current state only, and time is measured since entering the state.

The covariate $age04$ is included only for the hazard rate from state 0, therefore the coefficient $\beta_{04}(age)04 = .046 + .015 = .061$. The effect of age is decreased after a subject is diagnosed with either CVA or CHD. We found decreased effect of age only for the transition from CHD to death in males.

The role of diabetes changes depending on the current state. It is not significant for people without any disease, and has the highest effect for the hazard rate for $\alpha_{34}(t)$.

The effect of smoking is also increased after entering state 3. The hazard ratio prior to state 3 is $exp(.267) = 1.30$ and becomes $exp(.267 + .425) = 2.00$ after entering state 3.

Cholesterol is found to affect only the hazard rate 14.

Table 4.8: Mortality model , Females

Model for Males, transitions to CVA						
_t	$\hat{\beta}$	*Std. Err.*	*z*	*P>|z|*	*95 % Conf.*	*Interval*
age	.046486	.0053413	8.70	0.000	.0360172	.0569547
*age*04	.0148638	.0070911	2.10	0.036	.0009655	.0287621
psm	.2670933	.0596187	4.48	0.000	.1502428	.3839438
*psm*34	.4258012	.2248067	1.89	0.058	-.0148118	.8664143
spf	.0029533	.001124	2.63	0.009	.0007502	.0051564
*diab*14	.6133789	.2216572	2.77	0.006	.1789388	1.047819
*diab*24	1.057427	.4634109	2.28	0.022	.1491587	1.965696
*diab*34	1.220815	.5329612	2.29	0.022	.1762306	2.2654
*chol*14	.0021616	.0011919	1.81	0.070	-.0001745	.0044977

4.2.5 Estimating the cumulative hazard function

The models for males and females presented in Tables 4.8 and 4.6 have different baseline hazard rates depending on the transition rates. This makes it difficult to compare the actual hazard rates, so we will compare the cumulative hazard rates for different transitions. To illustrate this we consider a fixed set of values for the covariates are fixed as: total serum cholesterol=220, diabetes=1, smoking status=1, systolic blood pressure=130, age at entering the study= 52.

There are 4 cumulative hazard rates for females, each for the transition rates 04, 14, 24 and 34. The plotted functions represent the cumulative hazard, but evaluating the slope, we can compare the actual hazard rates.

The largest cumulative hazard is for the transition 34. It corresponds to a very high mortality rate shortly after entering state 3 and remains with highest slope. The cumulative hazard from state 2 is the second highest. The hazard rate is also very high initially and appears to have close to constant slope. A similar pattern is observed for the transition 14, however it is lower than from the CVA state. It appears that CVA alone is more fatal than CHD alone, for the specified set of covariates. The hazard from state 0 is very low at the beginning, increases slowly, until about 30 years, at which time it increases in an exponential

fashion. Since all subjects are initially at state 0, 30 years spent in the states, corresponds to 82 years of age. All the comparison here, is solely from the given state, i.e. if a female moves from state 1 to state 3, her hazard rate of dying is as the hazard rate from state 3, starting from $time = 0$.

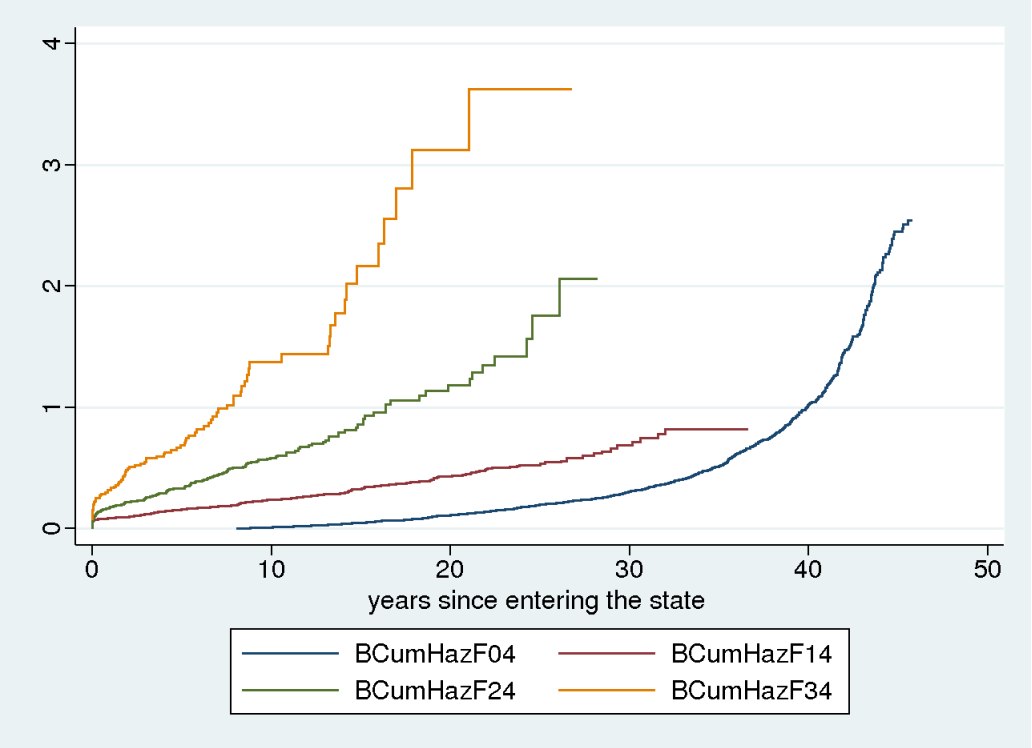

Figure 4.2: Cumulative hazards, Females

There are 5 cumulative hazard rates for males, the difference is that there are two hazard rates from state 3, depending on the order of the diseases, as captured by the covariate $cvafirst$. These two hazard rates are proportional, with a coefficient of proportionality $\exp(\beta_{cvafirst}) = 1.46$ higher for the individuals with a CVA diagnoses first. The cumulative hazards from state 1 and 2 are similar, with the curve for transition 14 being much smoother, since there many more transitions. If the cumulative hazard rates are plotted only for the initial period of time, we can observe that at the very beginning the hazard rate 14 is the highest. The cumulative hazard rate from state 0 has a similar shape to the female plot. The scale for the two plots is different, so any comparison has to take this into account.

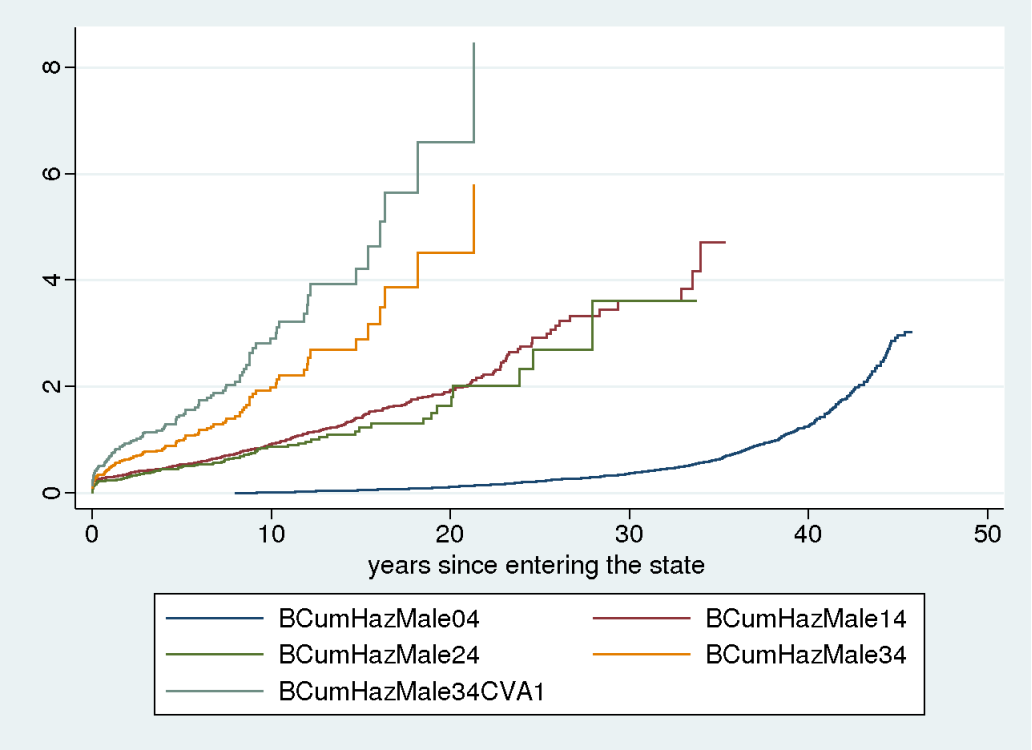

Figure 4.3: Cumulative hazards, Males

CHAPTER 5

Summary and Future Directions

5.1 Summary and Conclusions

We have applied multi-state models to the development and progression of CVD. Our goal was to develop prognostic functions predicting the probability of moving to a disease state (CHD or CVA) and the probability for death before and after developing these diseases. Another question, we examined is how different covariates affect the transition rates between various states. In Chapter 1 we discussed the currently available prognostic models for CVD. Chapter 2 outlined the necessary background from survival analysis and we introduced the Cox and the Aalen models. In Chapter 3 we compared the hazard rates for death before and after developing CHD. We used extended Markov models, since the mortality was increased immediately after the diagnoses. We also modeled and discussed how this initial peak changed with time. At this stage, we focused on evaluating the change of the mortality after CHD.

We estimated the hazard rates to mortality, allowing the risk factors to play different roles before and after the disease. We also modeled the hazard rate for developing CHD. The three transition rates were then combined in order to estimate the transition probabilities at time t, given the event history at time $s < t$. Transition probabilities can be used to evaluate individual prognoses for failure for a fix set of covariates.

The Aalen model was also considered in Chapter 3. The three hazards were modeled separately. The analysis from the Aalen model complemented the analysis from the Cox model. The first model has the ability to summarize the effect of a covariate and the Aalen model has the ability to show local changes (since the coefficients are estimated non-parametrically). Similarly, as in the Cox models, we estimated the transition probabilities.

This was done using the product integral introduced in Chapter 3.

Chapter 4 focuses on modeling mortality using a 5 state model, incorporating 2 more states, one for CVA and one for having both CHD and CVA. We have modeled the risk factors for developing CVA and whether CVA and CHD increase the likelihood of having the other disease. We compared different models and found that a semi-Markov model with a time scale, time since entering the study appears to explain the dynamics between the disease states. We estimated and compared the cumulative hazard rates for women with a fixed set of covariates.

The model for males was different, in the sense that the mortality rate from the state with both diseases depends on the order of the diagnoses. It is higher for individuals who developed CVA first, due to high mortality rate immediately after CHD. The models we developed can be used to distinguish the risk factors which are most important for mortality from each state.

The data set we used for this analysis is part of the Framingham Heart Study. The recruitment of the cohort took place between 1948 and 1952 in the city of Framingham, MA. Therefore any conclusion we derived from this data has potential limitations. First is that the population of the city of Framingham consist of predominantly white people and therefore the result may not apply to other races. Secondly, during the time the study took place the treatment of CHD improved noticeably, as improved health care and treatments became available. Also, we can assume the effect of some risk factors changed during this period. One example is blood pressure, with the availability of effective drugs for hypertension after the 1960's.

5.2 Future work

We have considered models with one and two disease states. To have a more complete picture of CVD progression, we can also include Congestive Heart Failure (CHF), which is a common disease after developing CHD. This however would increase the number of transitions, as adding one disease state in Chapter 4 resulted in 5 transitions. We used the Aalen additive model in Chapter 3 to compare with our results and to gain more insight into the role of the covariates. The additive models are easily adaptable for large number of states, since estimating the transition probabilities is obtained through the product formula, i.e. through

matrix multiplication. We have observed in Chapter 3, that using coefficient functions for a covariate has both advantages and disadvantages. An improvement in this direction could be using a semi-parametric submodel of the Aalen additive model, where the set of covariates is divided into two groups- the first group modeled non-parametrically (the coefficient for a covariate v is a function $\beta_v(t)$) and the second group is modeled parametrically (the coefficient is β_v, which corresponds to a constant effect of the covariate v). One such model was proposed by McKeague and Sasieni, [14]. As a submodel of the Aalen model, this model allows one to determine the variables whose effect varies with time. McKeague and Sasieni discussed the question of how to choose parametric or non-parametric effects. Further, Gandy and Jensen [29] developed formal tests for this model. Their results can be adjusted to detect particular alternatives, e.g. against the Cox model.

Flowgraph models were discussed in Chapter 2. They can be very useful in modeling multi-state models with a large number of states. Their use is limited for our application since they presently do not allow the use of covariates. One possible direction is extending these models to allow covariates, for example in the form of the accelerated time failure models. The estimated baseline hazard rates in Chapter 4 suggest that even models with constant baseline hazard (exponential density) or piecewise constant may be appropriate.

Another possible direction is to use multi-state models for recurrent events, for example repeated Myocardial Infarction. Frailty models are particularly useful in a model where the events are not independent. Bayesian analysis has been used in reliability and repair models. We are considering applying these techniques for evaluating the prognoses for individuals suffering multiple heart attacks.

APPENDIX A

Testing Proportionality Hazards assumptions

The Cox model assumes that the coefficients for all the covariates do not change with time. Grambsch and Therneau (Therneau, 1994) suggested a test for this assumption based on the scaled Shoenfeld residuals. The test is performed variable by variable and an overall test can also be calculated. Let assume there are no tied failure times, and fix a covariate x_u, $u = 1, \ldots, p$. Recall that the risk set R_j is defined to be the set of all the subjects alive at the time subject j fails. The Shoenfeld residual for the covariate X_u for the subject j observed to fail is:

$$r_{uj} = X_{uj} - \sum_{i \in R_j} X_{ui} c_{ui}$$

where c_{ui} are weights

$$c_{ui} = \frac{\exp(\boldsymbol{X_i \beta})}{\sum_{k \in R_j} \exp(\boldsymbol{X_k \beta})}$$

Let us consider the coefficient for the covariate X_u to depend on time and have the form:

$$\beta_u(t) = \beta_u + \gamma_j g(t)$$

where γ_j is a coefficient and $g(t)$ is a specific function of time. Grambsch and Therneau provide a method of scaling, to form r^*_{uj} so that

$$E(r^*_{uj}) = \gamma_j g(t)$$

This suggests that a plot of of r^*_{uj} versus t_j may be used to asses the proportional hazards assumption. Formal test procedures are based on a test statistic with $\chi^2(1)$ and can be combined for an overall test for all variables.

APPENDIX B

(Partial) Likelihood Ratio Test

There are 3 main tests about the regression parameters $\boldsymbol{\beta} = \beta_0, \ldots, \beta_p$ fo a Cox model:
-Wald test,
-Score test and
-(Partial) Likelihood ratio test.
Often, we want to test a hypothesis of the form:

$$H_0 : \boldsymbol{\beta_1} = \boldsymbol{\beta_{10}}$$

where the coefficient vector split into two vectors: $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$. This is the case when we want to compare a model (M_2) with covariates $\{X_0, X_1, X_2\}$ and a model (M_1) with covariates $\{X_0, X_1\}$. The two models are nested, in the sense that the covariates for model (M_1) are included in the model (M_2). To compare the two models, we can split the coefficient vector
$\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$ into $\boldsymbol{\beta_1} = \beta_2$ and $\boldsymbol{\beta_2} = (\beta_0, \beta_1)$
and perform the test $H_0 : \beta_2 = 0$. To test H_0 we can use the Likelihood Ratio Test statistic:

$$X_{LR} = 2(LL_2 - LL_1)$$

where LL_2 and LL_1 are the log of the maximum likelihoods calculated using models (M2) and (M1). The test static X_{LR} has an asymptotic $\chi^2(1)$ distribution under H_0. In general, if the models $(M2)$ and $(M1)$ differ in q covariates, the test statistic X_{LR} has an asymptotic $\chi^2(q)$.

REFERENCES

[1] T. Gordon, W. Kannel, and M. Halperin. Predictability of coronary heart disease. *Journal of Chron. Diseases*, 32:427–440, 1979. 1.1.1

[2] T. Troels, D. McGee, M. Davidsen, and T. Jorgensen. A cross-validation of risk scores for coronary heart disease mortality based on data from the glostrup population study and the framingham heart study. *International Journal of Epidemiology*, 31:817–822, 2002. 1.1.1

[3] D.R. Cox. Regression models and life tables (with discussion). *Journal of the RoyalStatistical Society, Series B*, 34:187–220, 1972. 1.1.2

[4] R. Wilson, P. D'Agostino, D. Levy, A. Belanger, H. Silbershatz, and W. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–1847, 1998. 1.1.2

[5] J. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, Hoboken, N.J., 2002. 1.1.4, 1.1.4

[6] K. Anderson. A nonproportional hazards weibull accelerated failure time regression model. *Biometrics*, 47:281 –288, 1991. 1.1.4

[7] K. Anderson, P. Wilson, P. Odell, and W. Kannel. An updated coronary risk profile. a statement for health professionals. *Journal of the American Heart Association*, 83:356 –362, 1991. 1.1.4, 3.3

[8] R.M. Conroy and et. al. Estimation of then-year risk of fatal cardiovascular disease in europe: the score project. *European Heart Journal*, 24:987 – 1003, 2003. 1.1.5

[9] E.L. Korn, GraubardB.L., and D. Midthune. Time to event analysis of longitudinal follow-up of a survey:choice of the time scale. *American Journal of Epidemiology*, 145:72–80, 1997. 1.1.5, 3.3, 4.2.3

[10] A. Peeters, A. Mamun, F. Willekens, and L. Bonneux. A life course analysis of the original framingham heart study cohort. *European Heart Journal*, 23:458–466, 2002. 1.1.5, 2.6.1, 2.6.2

[11] G. De Backer and et. al. European guidelines on cardiovascular disease prevention in clinical practice. *European Journal of Cardiovascular Prevention and Rehabilitation 2003,*, 10:S1–S10, 2003. 1.1.5

[12] O. Aalen. A linear regression model for the analysis of life tables. *Stat. Med.*, 8:907–925, 1989. 2.2.2, 3.4.3

[13] O. Aalen, O. Borgan, and H. Fekjer. Covariate adjustment of event histories estimated from markov chains: The additive approach. *Biometrics*, 57:993–1001, 2001. 2.2.2, 3.1, 3.4.3

[14] I. McKeague and P. Sasieni. A partly parametric additive risk model. *Biometrika*, 81:501–514, 1994. 2.2.2, 5.2

[15] M.J. Crowder. *Classical Competing Risks*. Chapman and Hall, 2001. 2.3

[16] O. Aalen and S. Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of statistics*, 5:141–150, 1978. 2.4.2

[17] S. Mason. Feedback theory: some propertiesof signal flow graphs. *Proc. of the Ins. of radio engineers*, 41:1144–1156, 1953. 2.5.3

[18] H.E. Daniels. Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650, 1954. 2.5.3

[19] O.E. Barnoff-Nielsen and D.R. Cox. *Asymptotic techniques for use in statistics*. Chapman and Hall, 1989. 2.5.3

[20] O.E. Barnoff-Nielsen. Comments on saddlepoint methods and statistical inference. *Stat. Science*, 3:279–229, 1988. 2.5.3

[21] N. Welton and A. Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, Nov-Dec:633–645, 2005. 2.5.5

[22] C. L. Chiang. *Introduction to Stochastic Processes in Biostatistics*. Wiley, Hoboken, N.J., 1968. 2.6

[23] P. Hougaard. Multi-state models: a review. *Lifetime data analysis*, 5:239–264, 1999. 3.1

[24] E. Fix and J. Neyman. A simple stochastic model of recovery, relapse and loss of patients. *Human Bilogy*, 23:205–241, 1951. 3.1

[25] E. Sverdup. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Scandinavis Aktuarietids skrift*, 48:184–211, 1965. 3.1

[26] J. Klein and M. Moeschberger. *Survival Analysis, Techniques for Censored and Truncated Data*. Springer-Verlag, 2003. 3.4.3, 3.5.1, 3.5.1

[27] O. Aalen. Further results on the non-parametric linear regression model in survival analysis. *Stat. Med.*, 12:1569–1588, 1993. 3.4.3

[28] P.K. Andersen and N. Keiding. Multi-state models for event history analysis. *Statistical Methods in medical research*, 11:91–115, 2002. 3.5.1

[29] A. Gandy and U. Jensen. checking a semi-parametric additive risk model. *Lifetime data analysis*, 11:451–472, 2005. 5.2

BIOGRAPHICAL SKETCH

Dimitre Stefanov

Dimitre Stefanov was born in the city of Varna, Bulgaria in 1970. He received his B.S. and M.S. degrees in Theoretical Mathematics from the Sofia University, Sofia, Bulgaria. His Masters thesis was in the area of Algebra, under the guidance of Prof. Vesselin Drensky. He worked for two years in the Bulgarian Academy of Sciences, Institute of Mathematics and Informatics. Dimitre also has M.A. degree in Applied Mathematics from the University of Southern California.