LUND
UNIVERSITY

The effect of behavioural changes over time on Cox proportional hazards estimates

Jonatan Hedberg

Master's thesis in Statistics (15 ECTS)
Supervisor: Björn Holmquist
March 2018

Abstract

This thesis considers bias when studying behaviours in a Cox proportional hazards model. In Cox proportional hazards regressions and cohort studies in general, measurements are often made during a limited period of time. Behaviours may, however, change rather dramatically over time, and if these changes are unknown, they will distort the results in the regression models. We study this problem in the context of the effects of smoking and physical activity on cardiovascular disease by simulating Cox proportional hazards models. Changes in behaviour are simulated with Markov chains in four scenarios. In each scenario we perform ten sets of simulations where each set has a different transition probability.

The first scenario considers a dichotomous variable indicating physical inactivity. We find that an increasing probability of changing behaviour will eventually completely dilute the baseline estimates. In the second, third, and fourth scenario we instead look at a smoking status variable containing the categories smoker, ex-smoker, and non-smoker. The three-category variable was in the regressions decomposed into the two dichotomous variables *Smoker* and *Ex-smoker*. In the second scenario we only allow transitions from smoker to ex-smoker. That leads to the hazard ratio estimates of *Smoker* going towards the hazard ratio of *Ex-smoker*. In the third scenario transitions are also allowed from ex-smoker to smoker. This results in the hazard ratios of the two variables moving towards each other, as the transition probabilities become larger. Lastly, the fourth scenario have Markov chains where non-smokers are additionally allowed to transition to smokers. There we find that the hazard ratios of *Smoker* and *Ex-smoker* go towards 1.0 when the transition probability of going from non-smoker to smoker is large.

1 Introduction

Cohort studies are frequently used to gain insights on the impact of behavioural factors on mortality or the occurrence of certain diseases such as cancer and cardiovascular diseases. Behaviours, in addition to an assortment of other measurements, are often measured at a certain point in time. The measurements then form a baseline that is used in a regression such as Cox proportional hazards regression. Behaviours may, however, change over time from one behaviour to another. Such changes will naturally have an impact on regression estimates if only the behaviours at baseline are known, and may lead to incorrect conclusions regarding the effect of different behaviours on the risk of diseases.

The notion of bias in epidemiological research has been touched upon by many authors. Armstrong [1] for example, provides an overview of the effect of different types of bias on classical regression and relative risk estimates. An account for measurement bias for normal covariates in multiple regression and logistic regression is given by Reeves, Cox, Darby, and Whitley [18] using mathematical arguments and simulations. Regarding Cox proportional hazards regressions, literature focuses on so-called "immortal-time bias". This bias may occur when one studies a treatment for a disease that occurs at a point in time. If the treatment occurs prior to baseline measurements but is treated as if it was made during or before the baseline, a treated patient will per definition survive between the baseline and the time of treatment. The patient will, thus, be immortal for a period of time, resulting in an over-estimation of the treatment effect. Austin, Mamdani, Walraven, and Tu [3] use Monte Carlo simulations of 108 different scenarios, and show that the time that the treatment occurs affects the size of the bias. More recently, Jones and Fowler [11], show that the bias is larger when the hazard is increasing over time than if it is decreasing. Beyersmann, Wolkewitz, and Schumacher [4], prove that the incorrect handling of treatments occurring over time will lead to an overestimation of treatment effect in the case when the coefficients of the Cox proportional hazards regression are estimated with the generalised rank estimate. They also show that the amount of censoring can have an effect on the bias. All of these studies deal with the situation where the time-dependent variable is dichotomous and only changes once throughout the study period. Therefore, the results of the studies tell us little of the bias that can be caused by unobserved changes in behaviour.

This thesis attempts to contribute to the knowledge of bias created by time-dependent covariates on baseline studies. More specifically the aim is to study how different rates of change in habits impact baseline hazard ratio estimates of the Cox's proportional hazards regression. This bias is evaluated in the context of smoking habits and exercising habits and their effect on cardiovascular diseases. The topic of this thesis was provided by Clinical Studies Sweden - Forum Syd.

To fulfil the aim of studying the effect of behavioural change, simulations are used. The simulations use discrete time Markov chains to simulate changes in behaviour. The times-to-event are simulated using truncated piece-wise exponential distributions through the algorithm developed by Hendry [10]. For

each simulation repetition, one proportional hazards regression is estimated using the baseline, and one proportional hazards regression is estimated using full information of the changes in behaviour. The estimates of the two regressions are then illustrated in order to evaluate the bias created by the unobserved changes.

This Master's thesis is structured as follows: first, in chapter 2, an outline is given on the Cox proportional hazards regression and discrete time Markov chains, along with an overview of the simulation of time-to-event with time-dependent covariates. Thereafter in chapter 3, the method of the simulations is described, starting with the simulation of the baseline and ending with the simulation of time-to-event. In chapter 4 the results of the simulations are presented, with explanations as to why these particular results were found. Lastly, in chapter 5 a discussion on the method and the results is provided.

2 Theory

This chapter gives a short overview of different concepts relevant for the analysis. The chapter mainly deals with concepts of survival analysis and Cox proportional hazards regression. In addition, the theory behind the method of simulating the time-to-event is presented, whereafter some concepts and terminology regarding Markov chains is provided.

2.1 Cox proportional hazards regression

The widely used semiparametric proportional hazards regression was first presented in 1972 by Cox [8], and has since then been widely applied in many fields. Its popularity stems primarily from its simplicity and the fact that no assumptions regarding the distribution of the time-to-event have to be made. Due to its popularity, the Cox proportional hazards regression can be found in most introductory text books in survival analysis. In the following paragraphs an overview of the proportional hazards regression will be provided. For more thorough descriptions of the Cox regression and survival analysis in general, see for example Klein and Moeschberger [12] and Collet [6].

Before providing an overview of the Cox proportional hazards regression regression, an account of some general definitions from survival analysis will be given. One of the most important concepts of survival analysis is the survival function. It is defined as:

$$S(t) = P(T > t) = 1 - F(t). \tag{1}$$

The survival function, thus describes the probability of an event not having occurred at time t. Another fundamental concept of survival analysis is the hazard function. The hazard function is defined as:

$$h(t) = \lim_{k \to 0} \frac{P(t \leq T < t + k | T \geq t)}{k}, \tag{2}$$

meaning that $h(t)k$ approximately measures the probability of an event occurring in the next instant, given that the event has not yet occurred. The hazard function is related to the survival function through:

$$h(t) = \frac{f(t)}{S(t)}. \tag{3}$$

In survival analysis the survival time or time-to-event is often assumed to be distributed according to the Weibull distribution with density function:

$$f(t) = \lambda\gamma t^{\gamma-1} \exp\left(-\lambda t^{\gamma}\right). \tag{4}$$

This gives the following hazard function:

$$h(t) = \lambda\gamma t^{\gamma-1}, \tag{5}$$

which allows for either an increasing ($\gamma > 1$) or decreasing ($\gamma < 1$) hazard over time. This property is useful when analysing the survival time of humans, as it often has an increasing hazard. Should $\gamma = 1$ the hazard becomes constant over time, and the time-to-event will be distributed according to the exponential distribution.

In the setting of the Cox proportional hazards regression we define for our observations $i = 1, ..., n$, $Y_i = \min(T_i, C_i)$, where T is the time to event and C is the time when an individual exits the study for any reason other than the event, i.e. becomes censored. We further define the variables $\delta_i, \mathbf{X}_i$, where δ_i is an indicator variable, indicating whether an observation is censored, and $\mathbf{X}_i$ a set of covariates for the ith individual. In the original regression model these covariates are constant, and are, thus, not allowed to change over time. This restriction is later dropped, but will for now be kept.

We let the hazard function at time t be:

$$h(t|\mathbf{X}) = h_0(t)\, c(\boldsymbol{\beta}^t \mathbf{X}), \tag{6}$$

where $h_0(t)$ is an arbitrary baseline hazard that is the same for all individuals at any particular point in time. $\boldsymbol{\beta}$ is a vector of parameters, and $c(\cdot)$ a pre-specified function. The arbitrary baseline hazard is in the Cox proportional hazards regression treated as a non-parametric component, while the vector $\boldsymbol{\beta}$ is treated as a parametric component. For that reason the model is referred to as semiparametric. The function $c(\boldsymbol{\beta}^t \mathbf{X})$ has to be positive, and is often conveniently defined as:

$$c(\boldsymbol{\beta}^t \mathbf{X}) = \exp(\boldsymbol{\beta}^t \mathbf{X}). \tag{7}$$

The name proportional hazards regression stems from an important result. If we look at two individuals with different covariate values $\boldsymbol{X}$ and $\boldsymbol{X^*}$, the ratio of their hazard rates (the hazard ratio) is given by:

$$\frac{h(t|\mathbf{X})}{h(t|\mathbf{X}^*)} = \frac{h_0(t) \exp\Big(\sum_{k=1}^{p} \beta_k X_k\Big)}{h_0(t) \exp\Big(\sum_{k=1}^{p} \beta_k X_k^*\Big)} = \exp\Bigg(\sum_{k=1}^{p} \beta_k (X_k - X_k^*)\Bigg), \tag{8}$$

i.e. a constant. The hazard rates are, thus, always proportional, which has given the regression model the name proportional hazards regression.

The parameters of the Cox regression model are estimated using maximum likelihood. We assume non-informative censoring , i.e. it occurs independent of $\boldsymbol{X}$, independent event times, and that there are no ties between the event times. We further, denote $t_1 < t_2 < ... < t_n$ as the ordered event times (censored and uncensored), and $R(t_i)$ as the set of individuals at risk prior to time t_i. The probability of a certain individual having an event at time t_i with covariate values $\boldsymbol{X}_i$ is given by:

$$\begin{aligned}
&P(\text{individual has event at } t_i | \text{one event at } t_i) \\
&= \frac{P(\text{individual has event at } t_i | \text{survival to } t_i)}{P(\text{one event at } t_i | \text{survival to } t_i)} \\
&= \frac{h(t_i|\boldsymbol{X}_i)}{\sum_{j \in R(t_i)} h(t_i|\boldsymbol{X}_j)} \\
&= \frac{h_0(t_i)\exp(\boldsymbol{\beta}^t \boldsymbol{X}_i)}{\sum_{j \in R(t_i)} h_0(t_i)\exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)} \\
&= \frac{\exp(\boldsymbol{\beta}^t \boldsymbol{X}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)}
\end{aligned}$$

As was shown above the baseline disappears as it is the same for all individuals at time t_i. Multiplying the probabilities over all events and only letting uncensored event times contribute, we get the following likelihood.

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^t \boldsymbol{X}_i)}{\sum_{j \in R} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)} \right)^{\delta_i} \tag{9}$$

This is called a partial likelihood as it does not contain the baseline hazard. The partial likelihood is, however, treated as a normal likelihood, and it has been shown to have similar properties [7] [21]. Also note, that the indicator variable δ_i will be zero at a censoring event. To estimate the parameters, the logarithm is taken on the partial likelihood, whereafter its maximum is found iteratively using for example the Newton-Raphson method.

The likelihood above assumed that are no ties between event times. In reality this is often not true, as times are usually recorded as intervals. There are several methods to accomodate for ties with the most widely known being Breslow's method [5] and Efron's method [9]. Efron's and Breslow's method are very similar when the number of ties are small. Efron's method is, however, closer to the correct partial likelihood based on a discrete proportional hazards model, and generally performs better when the number of ties is high [12].

Let d_i be the number of events at time t_i and τ_i be the set of individuals having the event at time t_i. Furthermore, let $\boldsymbol{S}_i$ be the sum of all covariate vectors over the individuals having the event at time t_i, i.e. $\boldsymbol{S}_i = \sum_{k \in \tau_i} \boldsymbol{X}_k$.

Breslow's likelihood is then given by:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^t \boldsymbol{S}_i)}{[\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)]^{d_i}} \right)^{\delta_i}. \tag{10}$$

Efron's likelihood is in turn given by:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^t \boldsymbol{S}_i)}{\prod_{k=1}^{d_i} [\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j) - \frac{k-1}{d_i} \sum_{j \in \tau_i} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)]} \right)^{\delta_i}. \tag{11}$$

As stated earlier, the Cox proportional hazards regression model can be extended to accomodate for time-dependent covariates. In the version with time-dependent covariates, we replace $\boldsymbol{X}$ with $\boldsymbol{X}(t)$ in the hazard function, so that we get:

$$\lambda(t|\mathbf{X}(t)) = \lambda_0(t) \exp(\boldsymbol{\beta}^t \mathbf{X}(t)). \tag{12}$$

This results in the following partial likelihood:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^t \boldsymbol{X}_i(t_i))}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j(t_i))} \right)^{\delta_i}, \tag{13}$$

where we once more assume non-informative censoring, independent event times, and no ties. Note that it is, also, assumed that the value of $X_j(t)$ is known for each time an individual is at risk. In reality, however, we only need to know the value of $X_j(t)$ for event times that are uncensored due to the way that the partial likelihood is formulated. In the presence of ties, methods such as Breslow's and Efron's likelihoods can be used as in the case with constant covariates.

2.2 Simulating a proportional hazards regression

Although not in abundance, there are methods for generating proportional hazards regression with time-dependent variables. Building upon the works of Leemis [13] and Leemis et al. [14], Austin [2] formulated closed form expressions for simulating time-to-event for three types of time-dependent variables when the time-to-event follows an exponential, Weibull, or Gompertz distribution. Austin's method, however, only allows for dichotomous or continuous time-dependent variables. Sylvestre and Abrahamowics [20] propose a permutational algorithm, where generated survival times are randomly matched with covariate values according to a predefined probability distribution. The strength of this algorithm lies in that it allows for any number of time-dependent covariates and that it does not require that the time-to-event follows a parametric distribution.

A rather elegant and more intuitive algorithm using a piece-wise exponential distribution was developed by Hendry [10]. The algorithm is itself an extension of the work of Zhou [22]. In Zhou's procedure we have a constant variable x_1 and a time-dependent variable $x_2(t)$ that switches from 0 to 1 at time t_1. We, furthermore, define a function $g(\cdot)$ that is a monotone increasing function where

$g(0) = 0$ and $g(t)^{-1}$ is differentiable. If W is a variable following a two-piece exponential distribution with rate:

$$\psi = \begin{cases} \exp(\beta_1 x_1) & \text{if } t \leq g^{-1}(t_1) \\ \exp(\beta_1 x_1 + \beta_2) & \text{otherwise} \end{cases} \tag{14}$$

then $g(W)$ will follow a Cox proportional hazards model with baseline hazard $h_0(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[g^{-1}(t)\right]$. Hendry [10] extends this concept to include an arbitrary number of time intervals and time-dependent (and constant) covariates using a truncated piece-wise exponential distribution. In Hendry's framework we first partition a time scale into J intervals $(0, s_1], (s_1, s_2], ..., (s_{J-1}, s_J]$. We then let a variable Z follow a piece-wise exponential distribution with density function:

$$\begin{aligned} k_Z(t) = \prod_{h=1}^{j-1} \exp(-\psi_h(g^{-1}(s_h) - g^{-1}(s_{h-1})))\psi_j \\ \exp(-\psi_j(t - g^{-1}(s_{j-1})))I\{g^{-1}(s_{j-1}) < t \leq g^{-1}(s_j)\}, \end{aligned} \tag{15}$$

for $j = 1, 2, ...J$. We further denote the corresponding distribution function as $K_Z(t)$. A variable W that is a truncated version of Z will then have the density function:

$$f_W(t) = \frac{k_Z(t)I\{g^{-1}(a) \leq t \leq g^{-1}(b)\}}{K_Z(g^{-1}(b)) - K_Z(g^{-1}(a))}, \tag{16}$$

where $[g^{-1}(a), g^{-1}(b)]$ is the support of W. As with Zhou's two-piece exponential distribution, $g(W)$ will now also follow a Cox proportional hazards model with baseline hazard $h_0(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[g^{-1}(t)\right]$, if $g(\cdot)$ is a monotone increasing function where $g(0) = 0$ and $g(t)^{-1}$ is differentiable. Time-dependent covariates are entered into the model through the rates of the piece-wise exponential distribution so that $\psi_j = \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)$. To simulate data using Hendry's framework the following steps are followed:

1. Define the function $g(\cdot)$.
2. Define a time scale and partition the time scale.
3. Define the bounds of truncation.
4. Define the number of observations.
5. Define the vector of coefficients $\boldsymbol{\beta}$.
6. For each observation:
   (a) generate a set of time-dependent covariates $\boldsymbol{X}$.
   (b) let $\psi_j = \exp(\boldsymbol{\beta}^t \boldsymbol{X}_j)$ for each time interval $j = 1, ..., J$.
   (c) generate W from a truncated piece-wise exponential distribution with ψ_j as the rate for the jth time period.
   (d) calculate the time-to-event $Y = g(W)$.

2.3 Discrete-time Markov chains

Since discrete-time Markov chains are used in this thesis to simulate changes in behaviours, some basic concepts and terminology are presented in the following paragraphs. A more detailed introduction to Markov chains can be found in Stirzaker [19].

We consider a sequence of random variables $(X_0, X_1, ...)$ that can take on values from a finite set S. The set S is called the state space, and $X_t = i$ is often referred to as X being in state i. For $X = (X_0, X_1, ...)$ to be a Markov chain the Markov property has to hold, meaning that:

$$P(X_t = i|X_0 = k, ..., X_{t-1} = j) = P(X_t = i|X_{t-1} = j), \tag{17}$$

for all $t \geq 1$ and all $i, j, k \in S$. To put differently, the probability of the variable X taking a certain value in the future only depends on the value that it had during the time period right before. If the probabilities do not change over time, i.e. if the chain is homogeneous, we can rewrite equation 17 to:

$$P(X_t = i|X_{t-1} = j) = p_{ij}, \tag{18}$$

where p_{ij} signifies the probability of transitioning from state j to state i.

The probabilities can be displayed in a matrix called a transition matrix. Assuming a simple three state Markov chain, with states i, j, and k the transition matrix would be as follows:

$$\begin{pmatrix} p_{ii} & p_{ji} & 1 - p_{ii} - p_{ji} \\ p_{ij} & p_{jj} & 1 - p_{ij} - p_{jj} \\ p_{ik} & p_{jk} & 1 - p_{ik} - p_{jk} \end{pmatrix}$$

Should $p_{ii} = 1$ the state i is called absorbing. Furthermore, we say that i is accessible from j if the chain in some way can reach state i if starting from state j. If j also is accessible from i the states are said to be mutually accessible. Another important concept in Markov chains is that of recurrence. If we define the first return to i if starting at i as:

$$T_i = \min\{n \geq 1 : X_t = i|X_0 = i\}, \tag{19}$$

then i is recurrent if $P(T_i < \infty) = 1$. If instead $P(T_i < \infty) < 1$ state i is called transient. Note that if defined as in equation 19, the first return does not require X to have had other states than i, however this is generally considered to be the case.

3 Method

This chapter will outline how the simulation study was performed. Throughout the simulations, estimates from a Cox regression when changes in the variables are known are compared to estimates based only on the baseline measurements.

Each regression contains the following five variables: *Sex, Age*, *Passive* (indicating physical inactivity), *Smoker*, and *Ex-smoker*. The corresponding coefficients can be found in table 1. The coefficients were decided upon through discussions with Clinical Studies Sweden - Forum Syd, except for the coefficient belonging to *Ex-smoker*. Efron's method is used, as it is better at handling larger amounts of ties. Time is in the simulations discrete with year-long time-periods. All sets of simulations, furthermore, use a sample size of 1000, and each simulation is repeated 1000 times. The following subsections will describe how the baseline is simulated, how the time-dependent variables are made, and how the time-to-event is generated. Lastly, four scenarios of simulations are presented.

Table 1: Variables used in the simulations with corresponding coefficients

Variable	Coefficient
Sex	log(1.0)
Age	log(1.5)/10
$Passive$	log(1.5)
$Smoker$	log(2.0)
$Ex\text{-}smoker$	log(1.2)

3.1 Simulating the baseline

For the purpose of realism, the baseline is simulated loosely based on the Malmö Diet and Cancer Study cohort [15]. The variables generated in the baseline are: *Sex*, *Age*, *Passive*, and *Smoking status*. The *Smoking status* variable is in the regressions transformed into two binary variables: *Smoker* and *Ex-smoker*. *Sex* is simulated using a binomial distribution with probability of 0.4 of being male. *Age* is drawn from a normal distribution with mean 60 and standard deviation 10. *Passive* is generated through a binomial distribution with probability of being inactive of 0.6 if male and 0.4 if female. Lastly, *Smoking status* is drawn from two multinomial distributions with probabilities for males being 0.28, 0.41, and 0.31 and for females being 0.26, 0.27, and 0.47 for being a smoker, ex-smoker, and non-smoker respectively. Note that only, *Passive* and *Smoking status* are treated as time-dependent rendering the other variables uninteresting for the analysis to follow. They are, nonetheless, kept for the simulations to closer emulate a study of cardiovascular diseases.

3.2 Simulating the time-dependence

The time-dependent variables are simulated using discrete time Markov chains, for each individual with initial states as given by the baseline. The transition matrices are furthermore set as constant and are the same for all individuals. While other ways of simulating changes between categories can be utilised Markov chains are simple and allows one to easily change probabilities of movements between states by changing the transition matrix. The Markov property

is, furthermore, most likely relatively reasonable when it comes to behavioural patterns.

Regarding the variable indicating physical inactivity, *Passive*, we have the states passive and active. The number of individuals being physically active could reasonable be assumed to be constant over time. We, therefore, set the probability of becoming active equal to the probability of becoming passive. If we let ξ be the transition probability of transitioning from passive to active or vice versa we get the transition matrix as follows:

$$\begin{pmatrix} 1-\xi & \xi \\ \xi & 1-\xi \end{pmatrix}$$

The Markov chain for smoking status has the states smoker, ex-smoker, and non-smoker. Naturally individuals are in the Markov chain not allowed to become non-smokers. In other words, the state non-smoker is inaccessible from both smoker and ex-smoker. Furthermore, an individual is not allowed to transition from non-smoker directly to ex-smoker. Letting γ be the probability of transitioning from the state smoker to ex-smoker during the time period, ζ the transition probability from ex-smoker to smoker and ψ the probability of becoming a smoker as a non-smoker we get the following transition matrix:

$$\begin{pmatrix} 1-\gamma & \gamma & 0 \\ \zeta & 1-\zeta & 0 \\ \psi & 0 & 1-\psi \end{pmatrix}$$

Throughout the simulations, we set $\gamma > \zeta, \psi$ to portray a higher chance of stopping smoking than starting smoking, which one can argue to be reasonable given the mean age of 60 in the simulations.

3.3 Generating time-to-event

When generating the time-to-event, the algorithm presented by Hendry [10] is used. The reasons for choosing this method are the fact that it is rather intuitive, due to the connection between a piece-wise exponential distribution and a proportional hazards model, and its ease of use because of some already existing R code. We assume a Weibull distribution for the time-to-event. This is in order to have an increasing hazard over time, making the simulations somewhat more realistic in comparison to a constant or decreasing hazard. The function $g(\cdot)$ used in the simulations will thus be:

$$g(t) = \lambda t^{\nu} \tag{20}$$

ν is set to 2, so that the hazard is indeed increasing. Furthermore, λ is set to 0.0002. The value for λ was decided upon as it gives somewhat reasonable survival times for the used models. The bounds of truncation for the truncated piece-wise exponential distribution are set to 0 and 60. A person will thus have

an event between 0 and 60 years from the baseline. The bound of 60 years can be seen as rather long, considering the mean age of 60. Hendry's method has, however, been shown to suffer from bias when the truncation bounds span short ranges [17]. Individuals are censored at the end of each simulation. The probability of an individual being censored is set to 0.4, and the indicator of censoring is added so that a person being censored happens at the time where the individual would have its event. The censoring is, thus, completely unrelated to the covariate values of each individual.

3.4 Four scenarios of simulation

The simulations are made in four scenarios, each studying different aspects of the model and consisting of ten sets of simulations. In the first scenario we look at the dichotomous variable *Passive*, and how changes in the variable after baseline affects the hazard ratio estimates. Different values of ξ in the transition matrix are used for each of the ten sets of simulations. The values are presented in Table 2, and are chosen to illustrate a step-by-step increase in transition probabilities, from relatively low to relatively high.

Table 2: Transition probabilities for scenario 1

Transition probabilities										
ξ	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20

The following three scenarios all study the smoking status variables. In the second scenario, we set the states ex-smoker and non-smoker as absorbing, meaning that individuals are only allowed to stop smoking. This leads to ζ and ψ being set to 0 for all simulations. γ on the other hand follows the same patterns as ξ, as can be seen in Table 3.

Table 3: Transition probabilities for scenario 2

Transition probabilities										
γ	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
ζ	0	0	0	0	0	0	0	0	0	0
ψ	0	0	0	0	0	0	0	0	0	0

In the third scenario we let the state smoker be recurrent so that ex-smokers can start smoking. Since it is probable that more individuals stop smoking than fall back into smoking, we keep $\gamma > \zeta$, leading to the values found in Table 4.

Table 4: Transition probabilities for scenario 3

Transition probabilities										
γ	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
ζ	0.01	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18
ψ	0	0	0	0	0	0	0	0	0	0

In the fourth scenario we attempt to evaluate the effect when changes are allowed from all states. Here ψ is set to equal to ζ as can be seen in Table 5 to allow for a rather high amount of noise coming from non-smokers starting to smoke.

Table 5: Transition probabilities for scenario 4

Transition probabilities										
γ	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
ζ	0.01	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18
ψ	0.01	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18

For computational purposes, the variable *Passive* is regarded as being constant for scenarios two to four, and *Smoking status* is regarded as constant for the first scenario. As changes occur independently in the different Markov chains, this should not have an effect on the result.

4 Results

In this chapter the results of the simulations are presented, along with explanations as to why these particular results were found. The results are structured according to the four scenarios starting with the first scenario.
The results of scenario one indicate as might have been expected that the effect of increasing rates of change over time in a binary variable dilutes the baseline estimate. This can clearly be seen in Figure 1 since the hazard ratio of the baseline estimate can be seen approaching 1.0 with increasing transition probability. We can also observe that the bias has almost completely negated the true hazard ratio when the transition probability is 0.16. This is a very high probability, nonetheless, it is clear that a significant bias is created already when the transition probability for changing state per year is 0.02. The dilution of the baseline estimate is caused by the baseline value of the variable *Passive* becoming less and less informative as the rate of change increases. In other words, an individual having $Passive = 1$ in the baseline will in the later simulations switch between $Passive = 1$ and $Passive = 0$, leading to the baseline value containing very little information about the time-to-event. The estimate using full knowledge remains unbiased, and seem to perform well throughout the simulations.

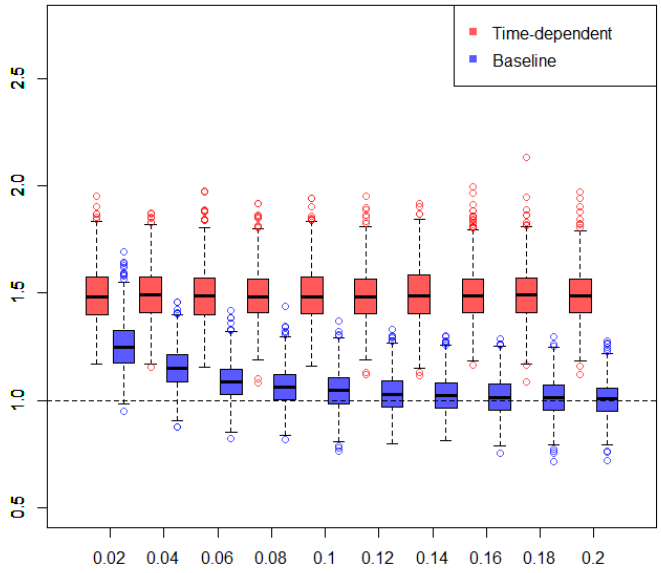
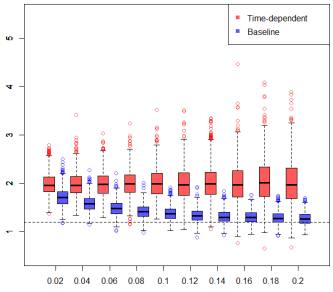
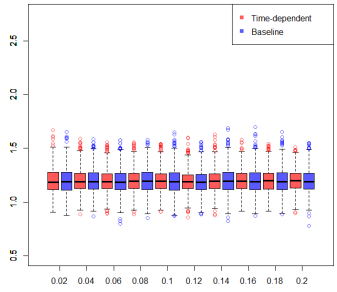


Figure 1: Scenario 1. The transition probability ξ is displayed at the bottom for each set of simulations. The y-axis indicates hazard ratio estimates with the dotted line indicating a hazard ratio of 1.0.

In the second scenario we instead look at the variables *Smoker* and *Ex-smoker*. We now, thus, have three categories instead of two. In this initial scenario individuals are only allowed to transition from *Smoker* to *Ex-smoker*. As the transition probability increases, more individuals who at baseline were smokers will now have stopped being smokers and become former smokers. They are in the baseline estimate, nonetheless, still incorrectly classified as smokers. For the baseline estimate, this means that more and more baseline smokers, will have times-to-event akin to that of baseline non-smokers as more baseline smokers stop smoking. The observed effect on time-to-event of baseline smokers will, therefore, appear to be closer to the effect of being a former smoker, as the difference between the two categories at baseline disappears. This gives us the results that can be found in Figure 2a, where the baseline estimate of the hazard ratio of *Smoker* approaches the hazard ratio of *Ex-smoker*. Although, the effect of smoking never disappears, it becomes indistinguishable from that of having stopped smoking at a probability of changing state of around 0.2. The estimates using time-dependent variables, are again unbiased, but show increasing variation with increased probability of change. Since none of the baseline former smokers change state, the estimates for *Ex-smoker* remain unchanged as seen in Figure 2b.



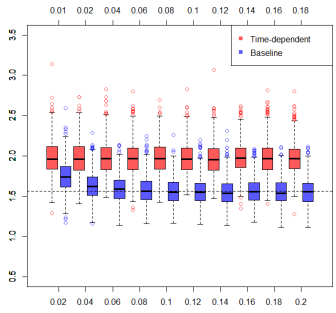
(a) Hazard ratios for *Smoker*.



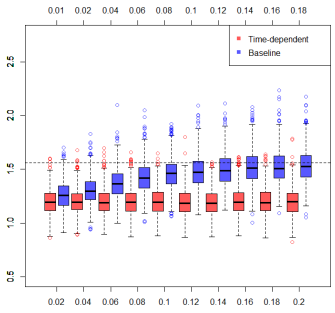
(b) Hazard ratios for *Ex-smoker*.

Figure 2: Scenario 2. The transition probability γ for each set of simulations is displayed at the bottom. The y-axis indicates the hazard ratio estimates. As neither ζ nor ψ are changed in this scenario, they are not displayed. The dotted line in Figure 2a indicates the hazard ratio 1.2, which is the same as the hazard ratio of *Ex-smoker*.

When letting the states ex-smoker and smoker be mutually accessible in scenario three, we get a different result. Individuals who at baseline are smokers may as before become former smokers. They can, however, now also return to smoking. In the same manner, former smokers at baseline may start smoking, and they may thereafter stop. This results in the difference between the two categories *Smoker* and *Ex-smoker* at baseline disappearing if many individuals transition frequently between the states. As in scenario two, this means that the perceived effect of *Smoker* will be more similar to the effect of *Ex-smoker*. However, in this scenario, the perceived effect of *Ex-smoker* will at the same time become more similar to that of *Smoker*. This leads to the results found in Figures 3a and 3b, i.e. that the two hazard ratios move towards each other with more changes in states. In essence what happens is that the difference in time-to-event between the categories *Smoker* and *Ex-smoker* at baseline disappears, and the only thing relevant at baseline is if an individual belongs to any of those two categories. Using the probabilities used for simulating the baseline values of *Sex* and *Smoking status* and the hazard ratios of the two categories, we can through simple probability calculations get that the expected weighted average hazard ratio of a combined class would be approximately 1.56. This is also the value that the hazard ratios seem to move towards. In the two graphs we can, furthermore, see that the hazard ratio of *Smoker* reaches 1.56 much quicker than *Ex-smoker*. To some extent, this caused by the probability of changing state always being higher for *Smoker*. Additionally, the probability of being a smoker at baseline is lower than that of being a former smoker. The baseline category *Smoker* will, thus, require fewer changes to become uninformative. As opposed to the previous scenario the variation in the estimates of the time-dependent variables is not dramatically affected.

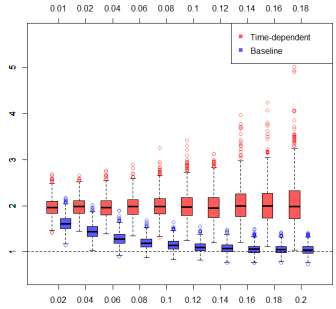


(a) Hazard ratios for *Smoker*.

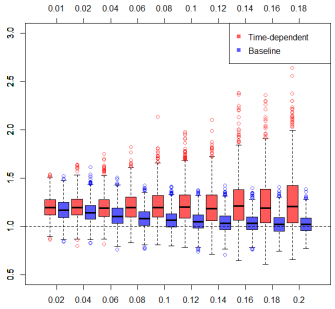


(b) Hazard ratios for *Ex-smoker*.

Figure 3: Scenario 3: γ is displayed at the bottom of each figure, while ζ is displayed at the top. The y-axis indicates hazard ratio estimates with the dotted line in the two figures indicating the hazard ratio 1.56.



(a) Hazard ratios for *Smoker*.



(b) Hazard ratios for *Ex-smoker*.

Figure 4: Scenario 4. The transition probability γ is displayed at the bottom. ζ is displayed at the top of the figures. ψ is throughout the scenario equal to ζ. The y-axis indicates the hazard ratio estimates.

In the fourth scenario the individuals who are non-smokers are also allowed to start smoking. This means that both *Smoker* and *Ex-smoker* will, after baseline, have an influx of baseline non-smokers. In the previous scenario where both smokers and former smokers could change state, the difference between the two categories at baseline disappeared. Due to non-smokers now also changing states, the difference between all categories at baseline becomes smaller when more individuals transition between the states. Therefore, as can be seen in Figures 4a and 4b, the effect of *Smoker* and *Ex-smoker* become incorporated into the baseline hazard. To put it differently, both hazard ratios go towards 1.0, and they seem to have done so at a probability of changing state of approximately

0.16 for *Smoker* and 0.18 for *Ex-smoker*. We can also note, that the variation in hazard ratio estimates increases rather dramatically for larger probabilities.

5 Discussion

The aim of this thesis was to study the effect of changes in behaviour after baseline measurements on the estimates of the Cox proportional hazards regression. The changes in behaviour were studied through simulations using Markov chains. The simulations have portrayed four scenarios, where in each we let the transition probabilities in the Markov chains have increasing values. In the first scenario, we looked at dichotomous variable indicating physical inactivity. There we found that a high transition probability leads to the baseline hazard ratio estimate going towards 1.0. This is caused by the baseline information containing less and less information relevant to the time-to-event. In the three following scenarios the three category smoking state variable were studied. In the first of these scenarios, states ex-smoker and non-smoker were absorbing states, meaning that individuals were only allowed to go from smoker to ex-smoker. This resulted in the baseline hazard ratio estimate of *Smoker* going towards the hazard ratio of *Ex-smoker*. The reason for this, is that the individuals being smokers at baseline, get times-to-event more alike that of individuals who at baseline are former smokers. At the same time, baseline values of *Ex-smoker* contain correct information as they do not change over time, which resulted in the hazard ratio estimates being correct. In the following scenario, former smokers were allowed to become smokers. From a baseline perspective, this meant that the effect on time-to-event for baseline smokers became the same as that for baseline former smokers. For that reason, the baseline hazard ratio estimates both went towards an expected weighted average hazard ratio. Finally, in the last scenario, non-smokers were also allowed to start smoking with the same probability as that of a former smoker. The effect of the baseline categories on the time-to-event, diminished because of the influx of non-smokers. This in turn led to both hazard ratio estimates going towards 1.0. We have, thus, through the simulations shown that unobserved changes in behaviour occurring after baseline measurements do have an impact on Cox proportional hazards measurements, and that the effect can become rather large if changes occur frequently.

One might argue that the fourth simulation is rather unrealistic as both non-smokers and former smokers have the same probability of starting smoking. The scenario illustrates, however, the effect of too much influx from a third category that in the regression is part of the baseline hazard. Should the transition probability of changing state from non-smoker to smoker be lower than it was in scenario four, one might expect a result that would lie in between the fourth and third scenario. We would, then most likely see that the hazard ratio estimates of *Smoker* and *Ex-smoker* approaching each other, while there at the same time would be a diluting contribution of non-smokers that would push the hazard ratio estimates down. We would, thus, probably have hazard ratio estimates

lower than those of scenario three. The hazard ratios would, nonetheless, not approach 1.0.

During the simulations it was noted that the variance of the estimates with time-dependent variables sometimes increased when the probability of changing state increased. This effect was most prominent in the last scenario, but could also be seen in scenario two. In scenario four, the population of non-smokers were allowed to start smoking, leading to a decrease in the number of individuals in this category since non-smoker was not accessible from any other state than itself. For simulations with a high probability of changing state from non-smoker to smoker, many observations will end up belonging to the states ex-smoker or smoker, which in the regressions are coded as two dichotomous variables. This means that the two variables will be highly correlated, leading to inflated coefficients and general instability in the estimates. This could be alleviated by simply changing the coding so that either non-smoker is coded as a variable in place of either smoker or ex-smoker. Regarding the second simulation, smokers are allowed to become former smokers, but not vice versa. If the transition probability is high, the category variable *Smoker* will eventually be filled with zeros, in turn resulting in few observed deaths with individuals belonging to smoker. This is likely the reason why we can see an increase in variance of the hazard ratio estimate for *Smoker* in scenario two, but not in scenario three. Naturally, problems with multicolinearity and coding of variables may arise when making any kind of cohort study. Extra caution should, nonetheless, probably be taken when performing cohort studies with time-dependent variables that change frequently.

One aspect that has not been dealt with in this thesis is censoring and the effect it may have on the bias of the baseline estimates. As censoring is modelled here, a censoring indicator is added at time-to-event, meaning that a censored individual is regarded as censored at the time were they normally would have their event. The individuals are, therefore, in the study for the same duration as if they would not be censored, and thereby is the number of changes in their smoking or exercising habits unaffected. It is, therefore, unlikely that the censoring affected the results in the simulation in any other way than by reducing the number of individuals with certain variable combinations having an event. A lower probability of censoring may, thus, have resulted in more stable estimates for scenarios two and four, but would otherwise have had no effect. Another type of censoring is censoring caused by a limited study time. To put differently, a study could run for a predefined or random number of years before the study is terminated and no more measurements are taken. The censoring time of an individual could also be generated from a distribution. In either case the time-to-event would be given by $Y_i = \min\{T_i, C_i\}$, where T_i is real the time-to-event and C_i the time to censoring. Should we have any of these two kinds of censoring, the length of the study may affect baseline estimates, since the amount of time that an individual is allowed to transition between states may be limited. A study with short times to censoring could, thus, potentially reduce the bias created by unknown changes in behaviour after baseline, especially when the transition probabilities are low.

The simulation of the effect of the different behaviours on time-to-event was in this thesis modelled in a simple way. The modelling of these effects, therefore, warrants some discussion. Throughout the simulations, we modelled the effect of behaviours as constant and instantaneous. When behaviours changed, the effect of past behaviours, furthermore, disappeared immediately. These assumptions are in many cases not realistic, in particular when it comes to smoking, since the positive effect of smoking cessation gradually appears. For example, a meta-study showed that former smokers in an elderly population have comparable risk of cardiovascular disease as that of non-smokers 20 years after smoking cessation [16]. This gradual effect could have been simulated as a set of dichotomous variables indicating different ranges of how long time that has passed since smoking cessation. The time-since smoking cessation could be simulated at baseline, and changed as both time passes and behaviours change. The baseline estimates of the category variables indicating time since cessation, would however be wrong by definition. Assume that one category would indicate time since cessation between 0 and 5 years. An individual belonging to this category at baseline, would as time passes, only belong to the category for maximum 5 years. Thereafter, the individual would belong to the next category, and so forth. Say that the individual has the event 10 years after baseline. The individual will now have contributed to the likelihood through events happening more than 5 years after baseline, while for those events, and possibly more, being incorrectly categorised as having stopped smoking 0 to 5 years ago. The only way that the categorisation will not be incorrect at some point is, therefore, if the individual has the event or is censored before the time since cessation has passed 5 years. This would, thus, always create a bias if the time-to-event is simulated using baseline measurements, even if no changes in smoking occur. Instead one would have to incorporate the inherent changes in categorisation caused by the passage of time. This could be achieved by also using a Cox proportional hazards regression with time-dependent covariates for the baseline estimates. Only the variables indicating time since smoking cessation would then change over time. One might, however, question if such a model is used to any greater extent in epidemiological research.

References

[1] Armstrong, B. G. "Effect of measurement error on epidemiological studies of environmetal and occupational exposures". In: *Occupational and environmental medicine* 55 (1998), pp. 651–656.

[2] Austin, P. C. "Generating survival times to simulate Cox proportional hazards models with time-varying covariates". In: *Statistics in Medicine* 31 (2012), pp. 3946–3598.

[3] Austin, P. C., Mamdani, M. M., Walraven, C. van, and Tu, J. V. "Quantifying the impact of survivor treatment bias in observational studies". In: *Journal of Evaluation in Clinical Practice* 12 (2006), pp. 601–612.

[4] Beyersmann, J., Wolkewitz, M., and Schumacher, M. "The impact of time-dependent bias in proportional hazards modelling". In: *Statistics in Medicine* 27 (2008), pp. 6439–6454.

[5] Breslow, N. E. "Analysis of Survival Data under the Proportional Hazards model". In: *International Statistics Review* 43 (1975), pp. 45–58.

[6] Collet, D. *Modelling Survival Data in Medical Research*. Boca Raton, US: Taylor & Francis Group, 2015.

[7] Cox, D.R. "Partial Likelihood". In: *Biometrika* 62 (1975), pp. 269–276.

[8] Cox, D.R. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220.

[9] Efron, B. "The Efficiency of Cox's Likelihood Function for Censored Data". In: *Journal of the American Statistical Association* 72 (1977), pp. 557–565.

[10] Hendry, D. J. "Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers". In: *Statistics in Medicine* 33 (2014), pp. 436–454.

[11] Jones, M. and Fowler, R. "Immortal time bias in observational studies of time-to-event outcomes". In: *Journal of Critical Care* 36 (2016), pp. 195–199.

[12] Klein, J. P. and Moeschberg, M. L. *Survival Analysis: Techniques for Cencored and Truncated Data*. 2nd ed. New York, USA: Springer Science+Business Media, Inc., 2003.

[13] Leemis, L. M. "Technical Note – Variate Generation for Accelerated Life and Proportional Hazards Models". In: *Operations Research* 35 (1987), pp. 892–894.

[14] Leemis, L. M., Shish, L., and Reynertson, K. "Variate Generation for Accelerated Life and Proportional Hazards Models with Time Dependent Covariates". In: *Statistics & Probability Letters* 10 (1990), pp. 335–339.

[15] Manjer, J., Carlsson, S., Elmståhl, S., Gullberg, B., Janzon, L., Lindström, M., Mattison, I., and Berglund, G. "The Malmo diet and cancer study: representativity, cancer incidence and mortality in participants and non-participants". In: *European Journal of Cancer Prevention* 10 (2001), pp. 489–499.

[16] Mons, U., Müezzenler, A., Gellert, C., Schöttker, B., Abnet, C., Bobak, M., Groot, L. de, Freedman, N. D., Jansen, E., Kee, F., Kromhout, D., Kuulasmaa, K., Laatikainen, T., O'Doherty, M. G., Bueno-de-Mesquita, B., Orfanos, P., Petters, A., Schouw, Y. T. van der, Wilsgaard, T., Wolk, A., Trichopoulou, A., Bofetta, P., and Brenner, H. "Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium". In: *BMJ* 350 (2015), pp. 1551–1563.

[17] Montez-Rath, M. E., Kapphahn, K., Marthur, M. B., Mitani, A. A., and Hendry, D. J. "Guidelines for Generating Right-Censored Outcomes from a Cox Model Extended to Accomodate Time-Varying Covariates". In: *Journal of Modern Applied Statistical Methods* 16 (2017), pp. 86–106.

[18] Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. "Some aspects of measurement error in explanatory variables for continuous and binary regression models". In: *Statistics in Medicine* 17 (1998), pp. 2157–2177.

[19] Stirzaker, D. *Stochastic Processes Models*. New York, US: Oxford University Press, 2005.

[20] Sylvestre, M. and Abrahamowics, M. "Comparison of algorithms to generate event times conditional on time-dependent covariates". In: *Statistics in Medicine* 27 (2008), pp. 2618–2634.

[21] Tsiatis, A. A. "A Large Sample Study of Cox's Regression Model". In: *The Annals of Statistics* 9 (1981), pp. 93–108.

[22] Zhou, M. "Understanding the Cox Regression Models With Time-Change Covariates". In: *The American Statistician* 55 (2012), pp. 153–155.