UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

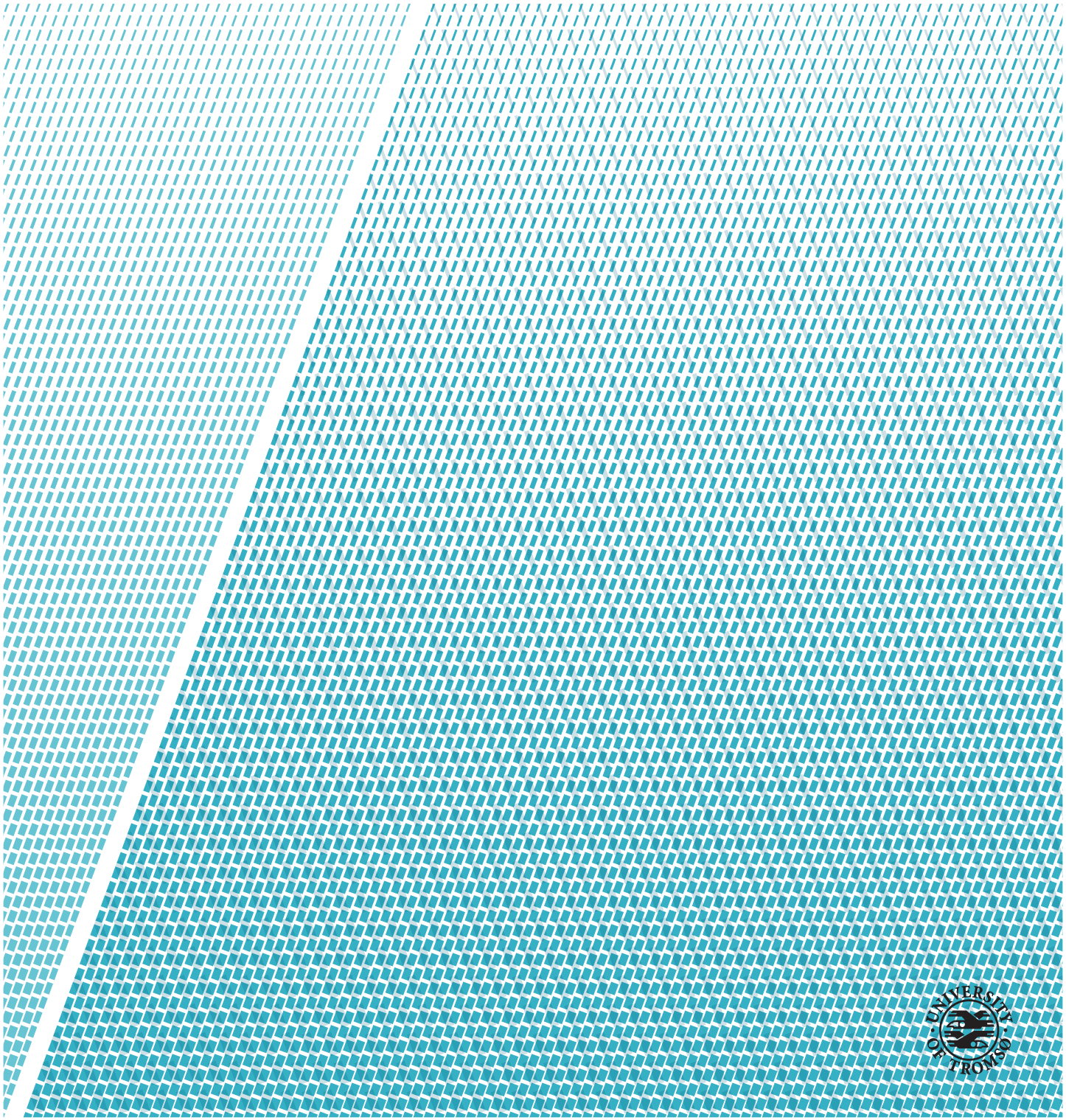
Faculty of Science and Technology
Department of Mathematics and Statistics

A study of the incidence and mortality hazard rate of myocardial infarction in Tromsø

Gunnhild Skjold

MAT-3907 Master's thesis in education - year 8-13, 40 SP

June 2019



Abstract

This thesis uses data from the Tromsø Study and weather data from the Norwegian Meteorological Institute to study the mortality hazard rate and incidence rate of myocardial infarction (MI) in Tromsø using a latent Gaussian modelling framework. Inference is performed using integrated nested Laplace approximations (INLA). This thesis presents the datasets and describes the modelling and computational framework, before analysis is performed.

To study the mortality hazard rate after MI, a Cox proportional hazards model has been implemented. A model without a seasonal effect with sex as a stratum variable was deemed the best fit. The results show an increased risk in the month after a MI. After the first month, the risk drops, before it increases with age. The mortality hazard rate is slightly higher for men than for women.

To study the change in the rate of MI during the time of the study, a Bayesian age-period-cohort model has been implemented. The model only includes the men of the study. This model studies the rate of MI on three different time scales: age, period, and cohort. The effects from age, period, and cohort are not directly identifiable. However, second differences describing the curvature and relative risk ratios are identifiable, as is the overall rate. The results show the incidence rate of MI decreasing with period, and increasing with age.

Acknowledgements

Eg vil sende ein stor takk til rettleiaren min Sigrunn Holbek Sørbye for at eg har fått sjansen til å skrive denne oppgåva. Takk for alle råd, den gode oppfølginga og all støtta! Ein ekstra takk for å ha klart å følgje meg opp så bra, sjølv om eg har sitte på andre sida av landet.

Eg vil takke alle vennane mine i Norsk Målungdom. Takk til Frida og Anna Sofie for at dokker har heldt ut å bu med meg gjennom heile masterarbeidet, sjølv når stresset var på topp. Takk til Fredrik og dokker andre som har lagt til rette for at eg har kunne tatt ein månad fri for å fullføre denne oppgåva.

Eg vil takke alle vennane mine frå studietida. Dokker veit kven dokker er. Ein spesiell takk går til Veronica: det har vore ein stor støtte å vere to lektorar som skriv fagleg master.

Til slutt: Takk til foreldra mine for å ha støtta meg gjennom skrivinga, og til mamma for å ha lese korrektur. Det har vore til stor hjelp å få bu hos dokker og bli varta opp når eg har vore heime i Tromsø. Eg håpar dokker synest det er stas med enda ein statistikar i familien!

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

According to the Norwegian Institute of Public Health [2009], cardiovascular diseases are the leading cause of death in Norway when looking at all age groups combined, and myocardial infarction (MI) and strokes cause one in four deaths on a world basis. Achieving a higher understanding of what leads to MIs is essential to prevent them.

There are several known risk factors for cardiovascular disease, such as smoking, diabetes, unhealthy diet, alcohol consumption, and low physical activity [Yusuf et al., 2004]. An association with weather has also been shown by, among others, Barnett et al. [2005], Auger et al. [2017], and Mohammad et al. [2018]. They have found links between the incidence rate of MI and temperature and snowfall. However, findings from Mohammad et al. [2018] and Hopstock [2012] indicate that the MI incidence rate in subarctic climates, such as the climate in Tromsø, is not as affected by weather. Using new methods and studying the data from another angle can serve to either falsify or confirm these previous results.

The Tromsø Study includes data from close to 40,000 participants from the municipality of Tromsø, over a time period of nearly 50 years. Weather data from the Norwegian Meteorological Institute is available from their website `http://eklima.met.no` and contains meteorological data from weather stations from all around Norway, including Tromsø. The Tromsø weather station has

been in operation since 1895, and records temperature, precipitation, snow depth, and wind data. This allows us to connect weather data to the incidences of MI from the Tromsø Study.

1.2 Aims and motivations of the thesis

The objective of this master thesis is threefold: To explore and study statistical methods, to study the mortality hazard rate after a MI, and to study the incidence rate of MIs in Tromsø. The hypothesis is that increased snowfall and lower temperatures will lead to an increase in the MI incidence rate. The incidence rate of MI is also expected to decrease during the time of the study. The incidence rate is expected to be higher for men than women, and for older age groups. These are known results from the Norwegian Institute of Public Health [2009]. The mortality hazard rate is expected to be higher after a MI, and increase with time as the participants age.

As this is a master's thesis in education, another aim is to explore a subject that is relevant for the Norwegian school system. In the new Norwegian curriculum, one of the new interdisciplinary subjects will be "Folkehelse og livsmeistring" ("Public health and life management skills") [Norwegian Ministry of Education and Research, 2017]. In addition, the use of statistical methods to study data is part of the mathematics curriculum [Norwegian Ministry of Education and Research, 2006, 2013]. Statistical methods are essential in public health research, and having an in-depth understanding of this type of research is highly relevant when teaching the subject.

1.3 Outline of the thesis

Chapter 2 describes the Tromsø Study dataset and the weather dataset, with introductory analysis of each dataset. Age and gender differences are studied, as well as the seasonal variation.

Chapter 3 presents the methodology used to analyse the datasets. The latent Gaussian modelling (LGM) framework and the inference method of integrated nested Laplace approximations (INLA) is presented. In addition, the chapter gives some useful prior models and model evaluation criteria.

Chapter 4 studies the hazard rate after a MI using a Cox proportional hazards (PH) model. This chapter presents the Cox PH model, and how it can be cast into a LGM framework. Several different model configurations are studied, and

the simplest model with the best fit is chosen. The results of the analysis are then presented and discussed.

Chapter 5 studies the incidence rate of MI during the time of the study using a Bayesian age-period-cohort (APC) model. The chapter presents the model itself, as well as the identification problem in APC analysis. The results of the analysis are presented and discussed.

Chapter 6 summarises the results from chapters 4 and 5, and suggests some further areas of research.

2

Datasets and introductory analysis

The data in this thesis comes from the Tromsø Study, a health study conducted in the municipality of Tromsø. To analyse seasonal variation, registrations of temperature and snow depth were collected from the Norwegian Meteorological Institute's website, `http://eklima.met.no`. In this chapter, the datasets are presented, and analysed using basic methods. The presented topics include the age and gender distribution of the dataset, the incidence rate of MI, and the seasonal variation of MI.

2.1 The Tromsø Study

2.1.1 The dataset

The Tromsø Study is a repeated population-based health study conducted in the municipality of Tromsø in Northern Norway. The study has been conducted seven times: in 1974, 1979-80, 1986-87, 1994-95, 2001-02, 2007-8 and 2015-16. Data collection was carried out by the Department of Community Medicine at UiT The Arctic University of Norway in collaboration with the Norwegian Institute of Public Health, the University Hospital of Northern Norway (UNN), and Tromsø City Council. A total of 40,051 different people have participated in

at least one of the surveys of the Tromsø Study. Of these, 18,510 participants have participated three or more times. Table 2.1 shows the number of participants in the different studies and the invited age groups.

Table 2.1: Examination year, age, sex, and number of attending subjects (*n*). Data are downloaded from `http://tromsostudy.com`.

Study wave	Examination years	Age (years)	*n*	Sex
Tromsø 1	1974	20-49	6,595	Men
Tromsø 2	1979-80	20-54	16,621	Men/women
Tromsø 3	1986-87	12-67	21,826	Men/women
Tromsø 4	1994-95	25-97	27,158	Men/women
Tromsø 5	2001-02	30-89	8,130	Men/women
Tromsø 6	2007-08	30-87	12,987	Men/women
Tromsø 7	2015-16	40-99	21,083	Men/women

The analyses in this thesis are based on data from Tromsø 1-6. The participants have been linked to data from the Norwegian Causes of Death Registry to record date of death. In addition, the date of MI has been recorded from admissions to UNN, the only hospital in the region. Independent endpoint committees have reviewed each case, ensuring that all cases of first-ever MI have been recorded, also when admitted to other hospitals [Jacobsen et al., 2012].

The dataset covers a total of 39,870 participants (19,896 men and 19,974 women). Of these participants, 4,248 (2,858 men and 1,390 women) have experienced at least one MI. The dataset contains information about the date of visit for each study, the participants' age at this date (age group for Tromsø 1), sex, date of emigration, date of death, and date of first MI.

2.1.2 Age of participants

The study includes participants born between 1897 and 1978. The mean age at first MI was 64.9 years, and the median age was 64 years, as shown in figure 2.1. In our dataset, Tromsø 1 only gives the age of the participants as a five-year age group (e.g. 30-34, 35-39, and so on). For participants who first participated in Tromsø 1, their age is set to the first year of their age group.

Figure 2.2 shows the mean age in each month of the total study. This time span is defined from the date of the first MI in the study (August 1962) to the last MI in the study (November 2014), including a total of 628 months. Participants have been defined as entering the study in January of the year they reach 30 years, and exiting the study in the month of MI or death.

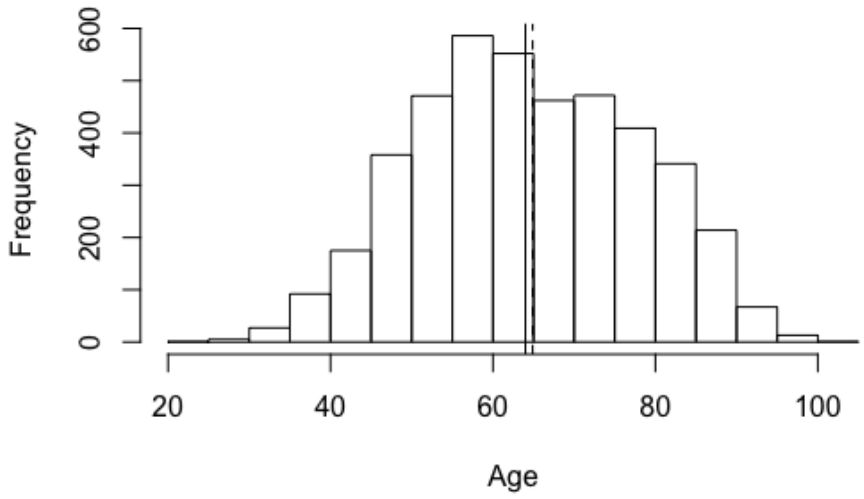


Figure 2.1: The distribution of age at first MI for the whole study sample. Mean age at first MI was 64.9 years (dashed line) and median age at first MI was 64 (solid line).

The mean age of participants has increased with time due to the different age groups invited for each study wave. Tromsø 4–6 invited older participants (as shown in table 2.1), and excluded the younger age groups, leading to an ageing study population.

Within each year, the mean age occasionally decreases. This is due to the number of active participants in each month. Participants are removed from the study upon MI or death, and most of these are older than the mean age. New participants only enter the study in January of each year. This effect is more pronounced after month 462, the month with the maximum number of participants. As there are few new participants each year and the active participants are older, the removal of older participants at death or MI has a larger effect on the mean age.

The mean age at MI has increased among the participants during the time of the study, from 44.1 years in the 1960s to 69.5 years in the 2010s. This is due to the increased number of older participants, and other methods are needed to study an eventual change in the mean age at MI.

Figure 2.3 shows the age distribution for first time participants. The age distribution at first participation is highly influenced by the invitation age span seen

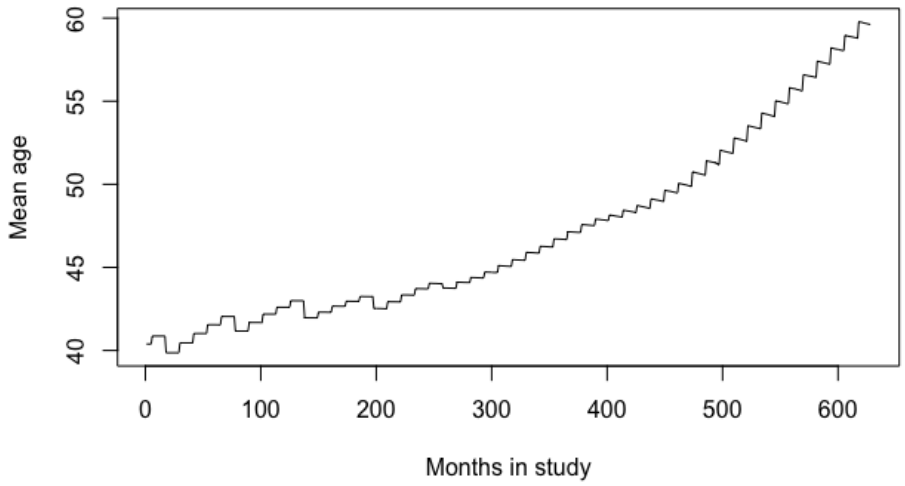


Figure 2.2: Mean age of participants in each month of the study, beginning with month 1 (August 1962) and ending with month 628 (November 2014).

in table 2.1. The figure shows a peak around 30 years, and most participants first participated in the Tromsø Study before they are 40 years. Combined with the ageing study population shown in figure 2.2, this shows that participants are participating in several Tromsø studies.

2.1.3 Gender

The number of men and women who have had a MI is shown in table 2.2. Approximately 16.8 % of men have had a MI, while 7.48% of women have had a MI. Men are expected to have a higher MI rate than women [Norwegian Institute of Public Health, 2009].

Table 2.2: MI incidences by gender

	Men	Women	Total
MI	2858	1390	4248
Not MI	17038	18584	35622
Total	19896	19974	39870

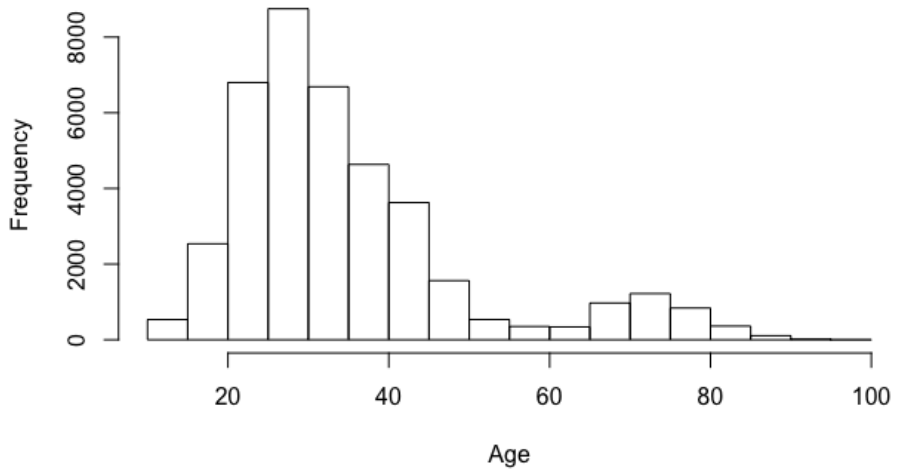


Figure 2.3: Age distribution at first participation.

To test this hypothesis in the Tromsø Study, a Pearson's chi-squared test is used to compare the number of men who get MIs with the number of women who get MIs. The observations are assumed to be independent. The test has test statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 573.47$$

$O_{i,j}$ are the counts for each combination of the the two variables, while $E_{i,j}$ are the corresponding expected values. The test shows that the number of participants who have had a MI is statistically significantly associated with gender with p-value $< 2.2 \cdot 10^{-16}$, and that men and women have different risks of having a MI.

2.1.4 MI rate and fatality ratio

According to the Norwegian Institute of Public Health [2009], the rate of mortality due to cardiovascular disease has decreased in the past 40 years. The rate of MI in the age group above 65 years has also decreased. However, this decrease is not seen in the younger age groups between 25–44 years.

In our dataset, the number of MIs per year increase during the time of the study. This is due to the increase both in the number of participants and their

age. Studying the rate of MI over time requires more advanced methods to separate the change in mean age and number of participants from the rate of MIs.

A simple correction for this is to look at the age-specific rates of MI per 1000 individuals, which are corrected for the number of participants in this age group at a certain time. Figure 2.4 shows the age-specific rate of MI per 1000 individuals during the time of the study. The age groups below 44 years and above 75 years have been excluded due to small sample size. There are missing years in some age groups due to the lack of MI in specific age groups at certain times. These rates are fairly constant for the younger age groups, while the rates for the older age groups increase at the start of the study and decrease at the end. A further study of the incidence rate will be performed in chapter 5.

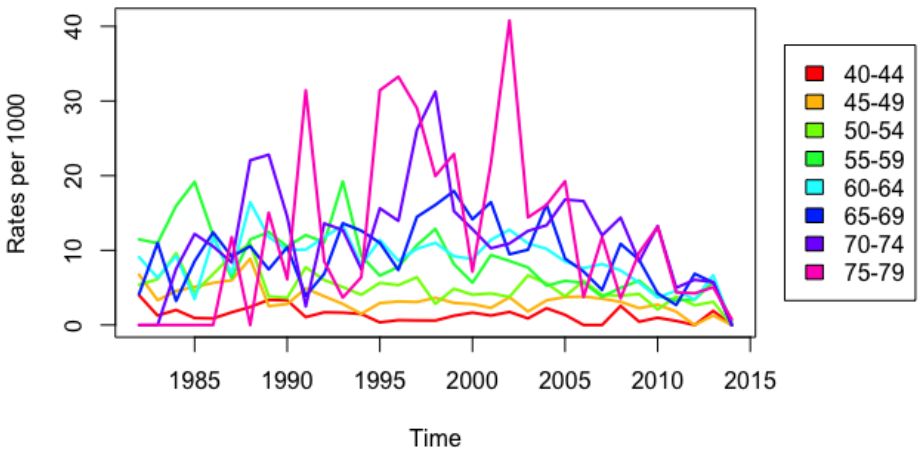


Figure 2.4: The rate of MI per 1000 for each age group over the time of the study.

About 12 % of deaths in the study are due to MI, and about 23 % of those who have had a MI die from it. Death by MI is defined as death occurring within 28 days of a MI [Hopstock, 2012]. As MIs cause a significant number of deaths in the study, studying the mortality hazard rate after MI and how it is affected by sex or season is important. A further analysis of the mortality hazard rate will be performed in chapter 4.

2.1.5 Seasonal variation

Hopstock [2012] showed that while 42 of 49 studies found an effect of temperature on MI, studies from the Nordic countries more often reported a lack of seasonal variation, including in the Tromsø Study. Hopstock [2012] reported that mean MI incidence was little affected by weather, but that winter weather (decreasing temperatures and increasing snowfall) led to an increased risk of MI in age groups above 65.

Figure 2.5 shows the mean number of MIs in each month of the year, adjusted for the number of days in each month, showing the seasonal variation. The figure shows a peak in the winter months November, December, and January. In these months, the mean number of MIs per month is above 12, and they clearly differ from the other months.

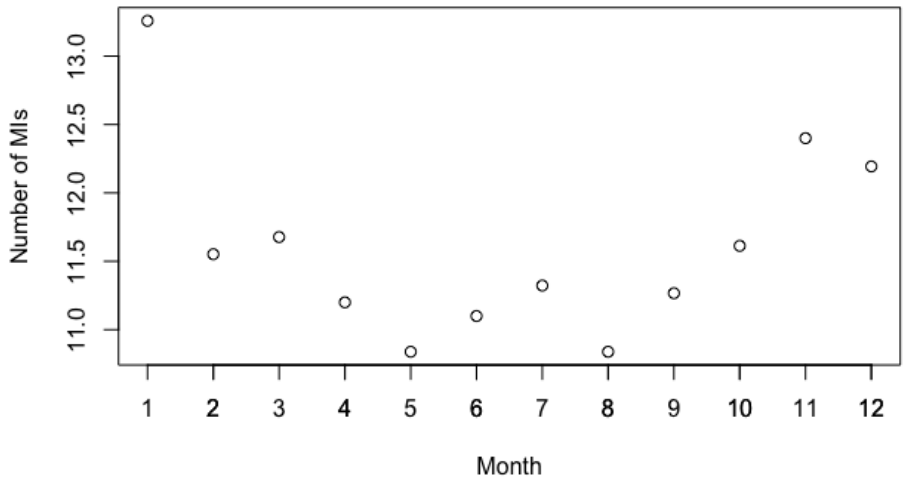


Figure 2.5: Seasonal variation for number of MIs per month, adjusted for the number of days in each month.

A two sample t-test can be used to study the difference in the mean number of MIs in the winter months and non-winter months. Winter months are defined as November, December, and January, with mean $\bar{x}_1$, sample variance s_1^2 and size n_1. The non-winter months have mean $\bar{x}_2$, sample variance s_2^2 and size n_2.

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{392 - 341.3333}{\sqrt{\frac{301}{3} + \frac{338}{9}}} = 4.31477$$

This shows that the mean number of MIs in the winter months November, December, and January is statistically significantly different from the number of MIs in the other months with a p-value of 0.015.

2.2 Weather data

2.2.1 The dataset

Weather data from the Norwegian Meteorological Institute is freely available from their website `http://eklima.met.no`. The data in this thesis is from the Tromsø observation station (station number 90450), and covers the time span from August 1962 to November 2014. This thesis makes use of observations of temperature (for each day, there has been recorded maximum, minimum, and mean temperature) and snow depth.

2.2.2 Temperature

Mohammad et al. [2018] studied the effect of air temperature with day-to-day incidence of MI in Sweden. They found a significant negative association with air temperature in all regions of Sweden, except in the north. This is consistent with the results shown by, among others, Hopstock [2012] and Barnett et al. [2005].

A simple linear regression model shows that the mean temperature of each year increases during the time of the study with a factor of approximately 0.03 ± 0.01 °C per year. If there are more MIs at cold temperatures, the number of MIs is expected to drop in recent years. However, the mean age of participants is higher in recent years, so this effect may be cancelled out.

Figure 2.6 shows the mean number of MIs per day at each temperature. Only temperatures with more than 40 MI incidents have been included. A simple linear regression model shows that there is a slightly positive relationship with temperature, which is not statistically significant with p-value 0.24. This is despite the seasonal variation that can be seen in figure 2.5. This variation has to be caused by other variables than temperature.

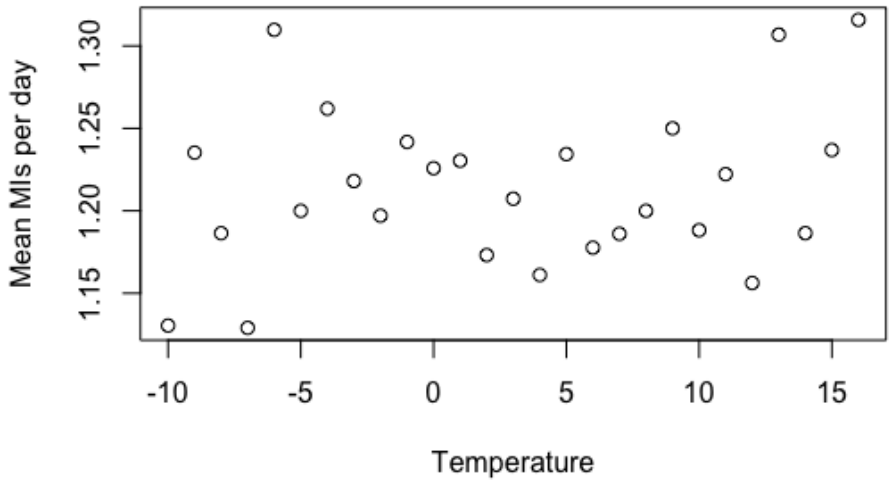


Figure 2.6: Mean number of MIs per day at each temperature.

2.2.3 Snow depth

One common myth is that shovelling snow increases the risk of a MI. News articles advising people to take care after large snowfalls can be seen in Norwegian media such as NRK (`https://www.nrk.no/norge/hjerteleger-advarer-mot-hard-snomaking-1.13923098`) and TV2 (`https://www.tv2.no/a/9691455/`). A connection between snowfall and risk of MI has also been shown by Auger et al. [2017], where a study of hospital admissions or death due to MI and snowfall in Quebec, Canada, showed that the risk of MI increased after snowfall among men.

We do not know which participants have shovelled snow, but we should see an effect on the number of MIs on days with large changes in snow depth if this myth is true.

Figure 2.7 shows the mean number of MIs per day in the following groups of snow depths (in cm): 1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, 91-100, 101-110, 111-120, 121-130, 131-140, 141-150, 151-160, 161-170, 171-180, and above 181. A simple linear regression model shows a slightly negative relationship that is not statistically significant with p-value 0.724.

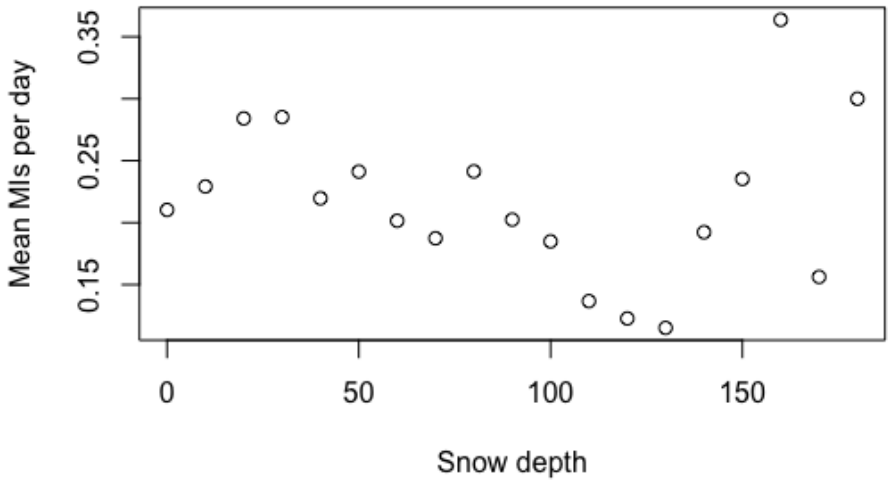


Figure 2.7: Mean number of MIs per day in each snow depth group.

Auger et al. [2017] suspect that men are more likely than women to shovel snow, and may be more exposed to a potential risk from increased snowfall. Therefore, the relationship between changes in snow depth and mean MIs per day is studied separately for each gender. Figure 2.8 shows the mean number of MIs per day for increases in snow depth between 1 and 20 centimetres for men and women.

Simple linear regression models for each gender show a slight negative relationship between snow depth change and mean number of MIs. These relationships are not statistically significant with p-values 0.415 (men) and 0.596 (women).

In this dataset, there is no observable connection between shovelling snow and having a MI, and the hypothesis that snowfall leads to an increased risk of MI does not have support in this study. However, our dataset does not state which participants have shovelled snow. A further study of this hypothesis would require more information.

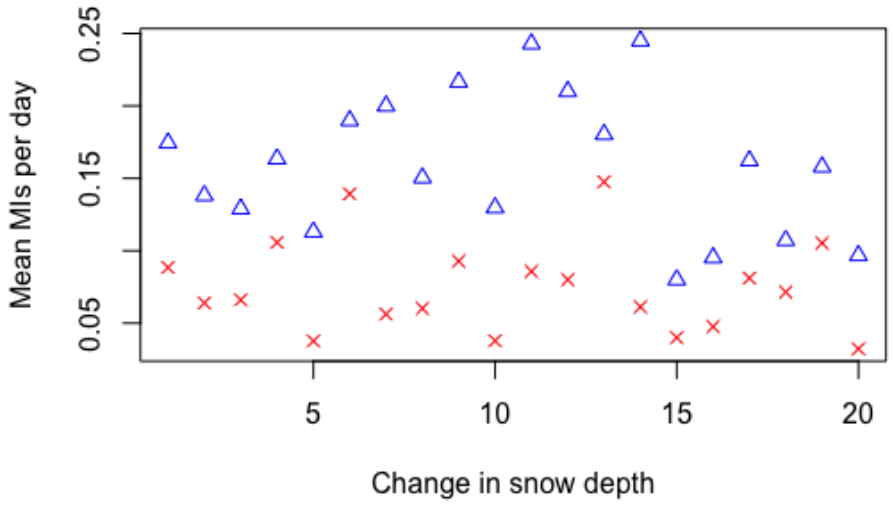


Figure 2.8: Mean number of MIs per day for change in snow depth, for men (blue dots) and women (red crosses).

3 Methodology

The last chapter showed that basic methods are not sufficient to analyse all the questions raised. The topics chosen for further study in this thesis are the mortality hazard rate and the MI incidence rate.

The methods used for analysis are presented in the following chapter. In this thesis a Bayesian framework is used, and inference will be performed using INLA, introduced by Rue et al. [2009]. INLA applies to the class of models known as LGMs. This chapter presents the modelling and the computational framework for inference for LGMs using INLA, as well as some relevant prior models and scoring rules.

3.1 Bayesian inference

A Bayesian framework is flexible, allowing us to account for information that is already known. Each parameter is viewed as a random variable, and inference is based on the prior beliefs about the variables in combination with observed data.

Bayes' rule for events is given as

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

It states that the conditional probability of A given B is dependent on the joint probability of A and B. In other words, the probability of A changes depending on the probability of B. When performing Bayesian inference, this rule is applied to probability distributions.

The prior distribution is a subjective probability representing our prior beliefs of how the parameters in the model will behave. The prior combined with the data contains all the information about the experiment. The posterior distribution is found by "updating" the prior with information from the experiment. The posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y} = [y_1, ..., y_n]$ is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{y})} \propto \pi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{3.1}$$

In this equation, $\pi(\boldsymbol{\theta})$ is the prior distribution, $g(\boldsymbol{y})$ is the marginal distribution of $\boldsymbol{y}$, and $\boldsymbol{y}$ is the data. When $y_1, ..., y_n$ are conditionally independent given $\boldsymbol{\theta}$, the likelihood is given as $p(\boldsymbol{y}|\ \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\ \boldsymbol{\theta})$.

In Bayesian inference, the goal is generally to find the posterior distribution of the parameters, given the data. From this posterior distribution, it is easy to calculate summary statistics, such as the posterior mean, median or mode.

Credible intervals can also be calculated using the posterior distribution. These intervals are similar to confidence intervals in frequentist statistics; however, while a confidence interval states that the interval will cover the true value of the parameter with a given probability, a credible interval states that the true value of the parameter will be within the interval with a given probability. A value is considered statistically significant when the credible interval does not cover zero.

For a long time, the issue with Bayesian inference has been how to perform the inference itself. By the time of the 90s and the early 2000s, Markov chain Monte Carlo (MCMC) methods had been developed, and computational frameworks such as `JAGS` and `BUGS` made Bayesian inference feasible. However, these methods are based on sampling, which can be time-consuming. In this thesis, the INLA methodology presented in section 3.3 will be used instead. This methodology uses approximation methods instead of sampling, and is therefore more computationally efficient.

3.2 Latent Gaussian models

Regression models are some of the most essential models in statistical analysis, and are used to study the relationship between variables or make predictions. The multiple linear regression model has predictor

$$\eta_i = E(Y_i) = \mu_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji}, \quad i = 1, ..., n \tag{3.2}$$

where $Y_1, ..., Y_n$ are independent random variables denoting the observations, α is the intercept, $\boldsymbol{z} = [z_1, ..., z_{n_\beta}]$ is a vector of covariates, and $\boldsymbol{\beta}$ is a vector of regression coefficients. This model only allows for linear effects, and the observations have a normal distribution $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

To study random variables $\boldsymbol{Y}$ that are not necessarily normally distributed, we can generalise the simple linear regression model. The expected values $\mu_i = E(Y_i)$ are linked to the linear predictors η_i with a link function $g(\cdot)$ so that $g(\mu_i) = \eta_i$ [Nelder and Wedderburn, 1972]. This gives a generalised linear regression model (GLM), where the random variables $\boldsymbol{Y}$ can have any distribution from the exponential family. This family of distributions includes, among others, the binomial, Poisson, and gamma distributions.

For this model, the linear predictor η_i has the same form as equation (3.2).

$$\eta_i = g(\mu_i) = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji}, \quad i = 1, ..., n \tag{3.3}$$

The GLM is a subclass of the general linear mixed model (GLMM), which can also include unstructured random effects ϵ_i, which are assumed to be independent and normally distributed with constant variance. The GLMM has the following predictor

$$\eta_i = g(\mu_i) = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \epsilon_i, \quad i = 1, ..., n \tag{3.4}$$

Further generalisations include generalised additive models (GAM) and generalised additive mixed models (GAMM), which allow for random or non-linear effects in the predictor. Specifically, the predictor of the GAM has the form

$$\eta_i = g(\mu_i) = \alpha + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}), \quad i = 1, ..., n \tag{3.5}$$

Here, the fs represent random or non-linear effects of covariates $\boldsymbol{u}$. The GAMM also includes unstructured effects ϵ_i. All these models can be expressed as

subclasses of a structured additive regression model [Fahrmeir and Tutz, 2001]. In this model, the structured additive predictor η_i has the form

$$\eta_i = g(\mu_i) = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \epsilon_i, \quad i = 1, ..., n \tag{3.6}$$

Here, the βs represent the fixed effects of the covariates $\boldsymbol{z}$, while the ϵs represent unstructured random effects. The functions f can describe many different effects, causing this framework to be very flexible. Examples of these effects can be temporally structured effects, spatially structured effects, random effects or non-linear effects of the covariates $\boldsymbol{u}$. It is clear that equations (3.2-3.5) can be expressed by equation (3.6).

In Bayesian analysis, the aim is to find the posterior marginals of all random quantities in the predictor η_i, shown in equation (3.6). To do this, a latent field $\boldsymbol{x}$ is defined. This field contains all the random variables of the linear predictor, in addition to the structured additive predictor $\boldsymbol{\eta} = [\eta_1, ..., \eta_n]$, so that

$$\boldsymbol{x} = [\boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \boldsymbol{f}(\cdot)] \tag{3.7}$$

The LGM is a special case of structured additive regression models where all the elements of the latent field are assigned Gaussian priors [Rue et al., 2009]. LGMs represent a unified computational framework containing several of the most common statistical models [Rue et al., 2009, 2017]. The models used in this thesis can be expressed within this framework, allowing us to use methods for inference applying to LGMs.

The LGM as a three-stage Bayesian hierarchical model has the following stages:

1. The first stage specifies the conditional distribution of the observations $\boldsymbol{y}| \boldsymbol{x}, \boldsymbol{\theta}_1 \sim \pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_1)$, where the dimension of $\boldsymbol{y}$ is n.

2. The second stage specifies the prior distribution of unobserved (latent) components $\boldsymbol{x}|\boldsymbol{\theta}_2 \sim \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_2)$. The dimension of the latent field is usually large, e.g. $n_x = 10^2$-10^5.

3. The third stage specifies prior beliefs about the hyperparameters $\boldsymbol{\theta}$ controlling the components in the model. The hyperparameters have distribution $\pi(\boldsymbol{\theta})$ and the dimension is often quite small, e.g. $n_\theta = 2$-5.

We assume that $\boldsymbol{x}$ has a multinormal distribution, $\boldsymbol{x} \sim \mathcal{N}(0, \Sigma)$. Observations $\boldsymbol{y}$ are assumed to be mutually conditionally independent, given the latent field $\boldsymbol{x}$ and the hyperparameters $\boldsymbol{\theta}_1$.

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_{i=1}^{n_y} \pi(y_i|x_i, \boldsymbol{\theta}_1)$$

The latent field is assumed to be a Gaussian Markov Random Field (GMRF). A GMRF is a multinormal random vector with Markov properties [Rue and Held, 2005]. This implies that the prior distribution of the latent field is defined by

$$\boldsymbol{x}|\ \boldsymbol{\theta}_2 \sim \mathcal{N}\big(0, \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)\big)$$

$\boldsymbol{Q} = \Sigma^{-1}$ is the precision matrix, the inverse of the covariance matrix Σ.

Precision matrices are commonly sparse. This is due to the fact that the precision matrix gives the structure of the conditional independence properties of the elements of $\boldsymbol{x}$. Specifically, the Markov properties of the GMRF state that $x_i \perp x_j|\boldsymbol{x}_{-ij} \Leftrightarrow Q_{ij} = 0$, where $\boldsymbol{x}_{-ij}$ represent all values of $\boldsymbol{x}$ apart from i and j. That is, x_i and x_j are conditionally independent, given the other values $\boldsymbol{x}_{-ij}$ [Rue et al., 2009].

For example, an auto-regressive process of order 1 with

$$\begin{aligned} x_1 &\sim \mathcal{N}(0, (1-\phi^2)^{-1}) \\ x_t|\ x_{t-1}, ..., x_1 &\sim \mathcal{N}(\phi x_{t-1}, 1), \quad t = 2, ..., n \end{aligned}$$

has hyperparameter ϕ and precision matrix

$$\boldsymbol{Q} = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & -\phi & & \\ & \ddots & \ddots & \ddots & \\ & & -\phi & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{pmatrix}$$

This is a sparse tridiagonal matrix, while the corresponding covariance matrix Σ is dense. Numerical methods for sparse matrices are far quicker than calculations for dense matrices, giving huge computational advances when using $\boldsymbol{Q}$ rather than Σ [Rue and Held, 2005].

The parameters of both the likelihood and the latent field are referred to as hyperparameters. The hyperparameters are not required to be Gaussian, and are denoted $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. They have distribution $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.

The joint posterior distribution for the latent field and hyperparameters is then summarised as

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \, \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i | \, x_i, \boldsymbol{\theta}) \tag{3.8}$$

This formulation represents an extension of the formulation in equation (3.1).

3.3 INLA

INLA is a computationally efficient method for Bayesian inference, requiring no sampling. MCMC methods apply to a wide range of models, while INLA only applies to LGMs, the type of model described in the previous section with predictor η_i, shown in equation (3.6).

With certain adjustments, the models in this thesis are LGMs, allowing us to use INLA to perform Bayesian inference. MCMC methods can also be used to perform inference on LGMs, but it is not well suited to these models due to the time required [Rue et al., 2009]. In addition, INLA is well suited to include random and non-linear effects in our models.

The aim is to find the posterior marginals for all components of $\boldsymbol{x}$ and all hyperparameters. The joint posterior of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ is shown in equation (3.8).

The target marginals for the hyperparameters is given by

$$\begin{aligned} \pi(\theta_j | \, \boldsymbol{y}) &= \int \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{\theta}_{-j}, \quad j = 1, ..., n_\theta \\ &= \int \pi(\boldsymbol{\theta} | \, \boldsymbol{y}) d\boldsymbol{\theta}_{-j} \end{aligned} \tag{3.9}$$

The target marginals for the components of the latent field is given by

$$\begin{aligned} \pi(x_i | \boldsymbol{y}) &= \int \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \, \boldsymbol{y}) d\boldsymbol{x}_{-i} d\boldsymbol{\theta}, \quad i = 1, ..., n \\ &= \int \pi(x_i, \boldsymbol{\theta} | \, \boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int \pi(x_i | \, \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} | \, \boldsymbol{y}) d\boldsymbol{\theta} \end{aligned} \tag{3.10}$$

INLA is performed in three steps [Rue et al., 2009, 2017]:

1. approximating the posterior marginal of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, by using the Laplace approximation,

2. computing an approximation of $\pi(x_i|\, \boldsymbol{y}, \boldsymbol{\theta})$ for selected values of $\boldsymbol{\theta}$,

3. using numerical integration and interpolation to combine the two previous steps to find an approximation to the target marginals.

3.3.1 Approximating the posterior of θ

To find the approximation to equation (3.9), $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is approximated by a Laplace approximation. The definition of conditional probability means that the posterior marginal of $\boldsymbol{x}$ can be written as

$$\pi(\boldsymbol{x}|\, \boldsymbol{\theta}, \boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\, \boldsymbol{y})}{\pi(\boldsymbol{\theta}|\, \boldsymbol{y})} \Leftrightarrow \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\, \boldsymbol{y})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}$$

The Laplace approximation is not used directly, as the estimated distribution of $\pi(\boldsymbol{\theta}|\, \boldsymbol{y})$ is typically not Gaussian. Therefore, the expression is rewritten. Tierney and Kadane [1986] showed that the Laplace approximation of a marginal posterior distribution can be written as

$$\tilde{\pi}_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}$$

where $\tilde{\pi}_G(\boldsymbol{x}|\, \boldsymbol{\theta}, \boldsymbol{y})$ denotes a Gaussian approximation evaluated at the mode $\boldsymbol{x}^*(\boldsymbol{\theta})$. This approximation method works well, as $\pi(\boldsymbol{x}|\, \boldsymbol{\theta}, \boldsymbol{y})$ is close to Gaussian in most cases. The distribution of $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\, \boldsymbol{y})$ is given in equation (3.8).

This gives the Laplace approximation for $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$

$$\tilde{\pi}_{LA}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|\, x_i, \boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{x}|\, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{3.11}$$

Note that the expression for $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ is originally given as

$$\begin{aligned}
\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) &\propto \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{y}, \boldsymbol{\theta}|\, \boldsymbol{x}) \\
&\propto \pi(\boldsymbol{x}|\, \boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|\, x_i, \boldsymbol{\theta}) \\
&\propto \exp\left\{ -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n} \log(\pi(y_i|x_i, \boldsymbol{\theta})) \right\}
\end{aligned}$$

The Gaussian approximation uses a Taylor expansion of the second order to approximate this expression.

$$\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*(\boldsymbol{\theta}))^T \boldsymbol{Q}^*(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{x}^*(\boldsymbol{\theta})) \right\} \tag{3.12}$$

Here, $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the location of the mode and $\boldsymbol{Q}^*(\boldsymbol{\theta}) = \boldsymbol{Q}(\boldsymbol{\theta}) + \text{diag}(\boldsymbol{c}(\boldsymbol{\theta}))$. $\boldsymbol{c}(\boldsymbol{\theta})$ is a vector containing the negative second derivatives of the log-likelihood with respect to x_i at $\boldsymbol{x}^*(\boldsymbol{\theta})$ [Rue et al., 2017].

3.3.2 Approximating the posterior of $\boldsymbol{x}$

The next step is to approximate the posterior marginal of $\boldsymbol{x}$, shown in equation (3.10). This requires approximations to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, which has already been found, and to $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$.

The approximation of the marginals of the latent field can be more challenging than approximating the marginals of the hyperparameters, as the dimension of $\boldsymbol{x}$ is assumed to be large. Using INLA, there are three options to estimate $\pi(x_i|\ \boldsymbol{\theta}, \boldsymbol{y})$, with varying speed and accuracy: 1) A Gaussian approximation, 2) a Laplace approximation, and 3) a simplified Laplace approximation.

The Gaussian approximation uses the GMRF-approximation shown in equation (3.12), calculated while approximating the posterior marginal of $\boldsymbol{\theta}$. However, this can be inaccurate, as this approximation assumes that the distribution is symmetrical, which is usually not the case. The Laplace approximation to $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ gives highly accurate results, but is computationally expensive.

Therefore, in this thesis, the simplified Laplace approximation (SLA) will be used to approximate the marginals of the latent field. The SLA is found by doing a series expansion of the Laplace approximation $\tilde{\pi}_{LA}(x_i|\ \boldsymbol{\theta}, \boldsymbol{y})$ around $x_i = \mu_i(\boldsymbol{\theta})$ and fitting it to a skew-normal density.

$$\log \tilde{\pi}_{SLA}(x_i|\ \boldsymbol{\theta}, \boldsymbol{y}) \propto bx_i - \frac{1}{2}x_i^2 + \frac{1}{6}dx_i^3 + \cdots$$

Then, b is a correction term for the mean and d is a correction term for skewness [Rue et al., 2017]. This method gives very accurate results, and is less computationally expensive than a full Laplace approximation.

3.3.3 Numerical integration

The approximated posterior marginals returned by INLA has the following forms [Martins et al., 2013]:

$$\tilde{\pi}(\theta_j|\,\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}\mid\boldsymbol{y})d\boldsymbol{\theta}_{-j} \tag{3.13}$$

$$\tilde{\pi}(x_i|\,\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\,\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\,\boldsymbol{y})\Delta\boldsymbol{\theta}_k \tag{3.14}$$

Rue et al. [2009] use $\tilde{\pi}(\boldsymbol{\theta}\mid\boldsymbol{y})$ to integrate out the uncertainty with regards to the hyperparameters when approximating equation (3.14). To do this, it is sufficient to have good evaluation points to perform the numerical integration. To find the integration points $\{\boldsymbol{\theta}_k\}$, the mode of $\tilde{\pi}(\boldsymbol{\theta}\mid\boldsymbol{y})$ is located. Then, new variables are constructed by using the negative Hessian matrix.

The Hessian matrix contains the second-order derivatives, and describes the curvature of the distribution. The Hessian can be used to find a reparametrisation that corrects for scale and rotation to explore the distribution and find relevant density values for integration [Rue et al., 2009]. $\tilde{\pi}(\boldsymbol{\theta}_k|\,\boldsymbol{y})$ are the density values computed during this exploration [Martins et al., 2013].

As $\boldsymbol{\theta}$ has a low dimension (n_θ = 2-5), it is possible and not too computationally expensive to derive the marginals for $\pi(\theta_j|\,\boldsymbol{y})$ from the approximation to $\boldsymbol{\theta}|\boldsymbol{y}$ using a grid exploration [Rue et al., 2017]. An integration free alternative to a grid-based approach is given in Martins et al. [2013].

3.4 Useful prior models

In this section, some useful prior models for the random effects in the LGMs are presented, being models for the $f(\cdot)$ effects in equation (3.6).

3.4.1 Random walk models of orders 1 and 2

Two common latent models are the random walk model of order 1 (RW1) and of order 2 (RW2), which are used to model non-linear trends and non-linear functions of covariates [Wang et al., 2018].

Rue and Held [2005] define the RW1 model with independent increments

$$\Delta x_i = x_i - x_{i-1} \sim \mathcal{N}(0,\tau^{-1}), \quad i = 1,\ldots,n-1 \tag{3.15}$$

with precision parameter $\tau = e^{\theta}$. The density is given by

$$\begin{aligned}\pi(\boldsymbol{x} \mid \tau) &\propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum(\Delta x_i)^2\right) \\ &= \tau^{(n-1)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}\right)\end{aligned}$$

Here, $\boldsymbol{Q} = \tau \boldsymbol{R}$ and $\boldsymbol{R}$ is a tridiagonal structure matrix, defined by Rue and Held [2005] as

$$\boldsymbol{R} = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}$$

Rue and Held [2005] define the RW2 model with independent second-increments

$$\Delta^2 x_i = x_i - 2x_{i+1} + x_{i+2} \sim \mathcal{N}(0, \tau^{-1}) \tag{3.16}$$

with precision parameter $\tau = e^{\theta}$. The density is given by

$$\begin{aligned}\pi(\boldsymbol{x}| \tau) &\propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum(\Delta^2 x_i)^2\right) \\ &= \tau^{(n-2)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}\right)\end{aligned}$$

As in the RW1 model, $\boldsymbol{Q} = \tau \boldsymbol{R}$ with structure matrix $\boldsymbol{R}$. Rue and Held [2005] give the structure matrix as

$$\boldsymbol{R} = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}$$

In both cases, τ can be seen as a smoothing parameter which is assigned a prior. The RW2 model smooths the effects more than the RW1 model.

In a LGM, $\boldsymbol{x}$ will represent one of the functions $f(\cdot)$, and $\tau = e^{\theta}$ will be a hyperparameter. The prior is defined on θ.

3.4.2 Independent random noise model

Another common prior model is the independent random noise model, where random variables $\boldsymbol{x}$ are assumed to be independent and identically distributed (iid) with density

$$\pi(\boldsymbol{x} \mid \tau) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \sqrt{\tau} \exp\left(\frac{1}{2}\tau x_i^2\right) \tag{3.17}$$

The precision parameter is $\tau = e^{\theta}$, and the prior is defined on θ.

3.5 PC priors

The chosen prior distribution of the hyperparameters might influence the result. However, choosing a prior can be difficult. Rue et al. [2017] describe several challenges surrounding the practice of choosing priors, and how priors often have been chosen due to computational convenience or because they are common in literature. Simpson et al. [2017] introduced a new class of priors, the penalised complexity (PC) priors. These priors are weakly informative, invariant to transformations, and penalise deviation from a base model.

Simpson et al. [2017] give the PC prior for the precision parameter τ of the models in section 3.4 as the distribution

$$\pi(\tau) = \frac{\kappa}{2}\tau^{-3/2} \exp\left(-\kappa\tau^{-1/2}\right), \quad \tau > 0, \kappa > 0 \tag{3.18}$$

where κ is a parameter indicating the penalty for deviating from the base model. This is easily transformed to give a prior for θ. A more detailed description of PC priors can be found in Simpson et al. [2017].

3.6 R-INLA and the BAPC package

The INLA methodology has been implemented in `R`, using the package `R-INLA`. The software is available from `http://r-inla.org`. The Bayesian APC (BAPC) model used in chapter 5 has also been implemented in the `R` package `BAPC`, available from `https://rdrr.io/rforge/BAPC/man/BAPC.html`.

3.7 Model evaluation criteria

3.7.1 Deviance information criterion

The deviance information criterion (DIC) is used to compare Bayesian models. It is based on adequacy and complexity. Deviance is a measure of fit or adequacy, and it is defined by Spiegelhalter et al. [2002] as

$$D(\theta) = -2\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$$

Then, Spiegelhalter et al. [2002] define the DIC as

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \tag{3.19}$$

where $D(\bar{\theta})$ is the posterior mean of the deviance, $\bar{D}$ is the posterior expectation of the deviance, and p_D is the effective number of parameters.

Lower values of the DIC indicate a better fitted model. It is known that the DIC may underpenalise complex models with many random effects [Plummer, 2008].

3.7.2 Logarithmic score

The logarithmic score (LS) was proposed by Good [1952] and can be given by the conditional predictive ordinate (CPO) as

$$\text{LS}_i = -\log(\pi(y_i \mid \boldsymbol{y}_{-i})) = -\log(\text{CPO}_i) \tag{3.20}$$

Gneiting and Raftery [2007] propose to look at the mean LS, $\overline{\text{LS}}_i = \frac{1}{n}\sum_{i=1}^{n}\text{LS}_i$. Lower values of the LS indicate a better fit.

Pettit [1990] defines the CPO as

$$\text{CPO}_i = p(y_i|\,\boldsymbol{y}_{-i})$$

where $\boldsymbol{y}_{-i}$ denotes all the observations except for y_i and $p(y_i|\,\boldsymbol{y_{-i}})$ is the predictive distribution. The CPO gives a measure of the probability of measuring a value.

Rue et al. [2009] give the predictive distribution for a LGM as

$$\pi(y_i|\,\boldsymbol{y}_{-i}, \boldsymbol{\theta}) = \int \pi(y_i|\,x_i, \boldsymbol{\theta})\pi(x_i|\,\boldsymbol{y}_{-i}, \boldsymbol{\theta})dx_i$$

The CPO is given by Held et al. [2010] as

$$\text{CPO}_i = \int \pi(y_i|\,\boldsymbol{y}_{-i}, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})d\boldsymbol{\theta} \tag{3.21}$$

The effect of $\boldsymbol{\theta}$ is then integrated out analogously as from equation (3.14), using numerical integration. However, the accuracy of the numerical integration depends on the approximation of $\tilde{\pi}(x_i|\, \boldsymbol{y}, \boldsymbol{\theta})$, which is required to approximate equation (3.21). This can be corrected for by manually computing new CPO values for the failed values [Held et al., 2010]. This is a built-in feature in the `R-INLA` package.

4

The Cox proportional hazards model

MIs are one of the leading causes of death worldwide and in Norway. A study of the mortality hazard rate and the factors influencing this rate can help to identify groups at risk. The mortality hazard rate represents the risk of death after having a MI. The Norwegian Institute of Public Health [2009] state that men are more at risk and that the risk increases with age, and this statement will be studied. In addition, we want to study a potential effect of season on the mortality hazard rate. To that end, a Cox PH model has been implemented. In this chapter, some relevant survival analysis concepts are introduced, the Cox PH model and its applications are described, and the results are presented. The Cox PH model is used to study the mortality hazard rate after a MI, and potential effects of sex, age, and season.

4.1 Survival analysis

The Cox PH model is one example of a survival model, which is utilised to analyse the time until an event. The distribution $F(t)$ of the survival time T describes the probability that a participant has died before time t, while the survival function $S(t)$ describes the probability of a random participant

surviving until time t. They are given by Kaplan and Meier [1958] as

$$\begin{aligned} F(t) &= P(T \leq t), \quad t \geq 0 \\ S(t) &= 1 - F(t) = P(T > t) \end{aligned} \tag{4.1}$$

T has density function $f(t) = \frac{dF(t)}{dt}$. When T is a continuous random variable, this also implies that $f(t) = -\frac{dS(t)}{dt}$.

The hazard function $h(t)$ gives the probability of experiencing the event at time t, given that the participant is alive at time t. Cox [1972] gives this hazard function as

$$h(t) = \lim_{s \to 0} \frac{P(t \leq T \leq t + s | \, T \geq t)}{s} \Rightarrow h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) \tag{4.2}$$

The cumulative hazard function $H(t)$ is then

$$H(t) = \int_0^t h(u)du = -\log S(t) \tag{4.3}$$

4.2 The Cox PH model

The Cox PH model was introduced by Cox [1972], and has since become one of the most common models for survival analysis [Wang et al., 2018].

It assumes the following hazard rate $h_i(t)$ for individual i with covariates $\boldsymbol{z}$.

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}_i \boldsymbol{z}_i), \quad i = 1, ..., n \tag{4.4}$$

$h_0(t)$ is a baseline hazard as a function of time and the predictor for an individual i is $\boldsymbol{\beta}_i \boldsymbol{z}_i$.

The ratio between two subjects a and b are constant in time, as $h_0(t)$ cancels out [Cox, 1972]. This is known as the proportionality assumption.

$$\frac{h_a(t)}{h_b(t)} = \frac{h_0(t) \exp(\boldsymbol{\beta}_a \boldsymbol{z}_a)}{h_0(t) \exp(\boldsymbol{\beta}_b \boldsymbol{z}_b)} = \frac{\exp(\boldsymbol{\beta}_a \boldsymbol{z}_a)}{\exp(\boldsymbol{\beta}_b \boldsymbol{z}_b)}$$

This basic Cox PH model only allows for linear effects. To increase utility, Martino et al. [2011] construct a piecewise log-constant proportional hazard model. This model is semi-parametric, and assumes a finite partition of the time axis $0 = s_0 < s_1 < \cdots < s_K$ with constant baseline hazard λ_k in each time interval. The baseline hazard is given as

$$h_0(t) = \lambda_k = \exp(b_k) \text{ for } t \in (s_{k-1}, s_k], \quad k = 1, ..., K.$$

We have n_β covariates $\boldsymbol{z}$ and n_f covariates $\boldsymbol{u}$. An individual i has hazard rate $h_i(t)$, where η_i is the predictor shown in equation (3.6), i.e.

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\eta_i) \\ &= \exp\left(b_k + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{m=1}^{n_\beta} \beta_{mi} x_{mi}\right) \\ &= \exp(\eta_{ik}), \quad t \in (s_{k-1}, s_k] \end{aligned}$$

Gaussian priors with unknown precision τ_b are assigned to the piecewise constant hazard $\boldsymbol{b} = [b_1, ..., b_K]$ and to the functions $\boldsymbol{f}$. Then, the predictors η_i are also Gaussian, and $\boldsymbol{x} = [\boldsymbol{\eta}, \boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{f}]$ is a GMRF.

The log-likelihood contribution of the ith observation at data point (t, δ), where t is the follow-up time and δ is an indicator variable stating whether death has occurred, is given by Martino et al. [2011] as

$$l = \log f(t) = \log(h(t)^\delta S(t))$$

As given in equation (4.2), $f(t) = h(t)S(t)$. For censored data, we include the indicator variable δ, which is 0 if the time has been censored and 1 if the event has occurred. Equation (4.1) gives the value of $S(t)$. Then, the log-likelihood contribution is

$$\begin{aligned} l &= \delta \log h(t) - \int_0^t h(u)du \\ &= \delta \log(\exp(\eta_k)) - \int_0^t \exp(\eta_k)du \end{aligned} \tag{4.5}$$

The time axis is partitioned so that $0 = s_0 < s_1 < \cdots < s_K$ and $t \in (s_{k-1}, s_k]$, $k = 1, ..., K$. The time t is not defined for time 0, so the integral in equation (4.5) goes from s_1 to t. Then, this expression can be rewritten as

$$\begin{aligned} l &= \delta\eta_k - \int_{s_k}^t \exp(\eta_k)du - \int_{s_{k-1}}^{s_k} \exp(\eta_{k-1})du - \cdots - \int_{s_1}^{s_2} \exp(\eta_1)du \\ &= \delta\eta_k - (t - s_k)\exp(\eta_k) - \sum_{j=1}^{k-1}(s_{j+1} - s_j)\exp(\eta_j) \end{aligned} \tag{4.6}$$

As this log-likelihood contribution depends on $\boldsymbol{\eta}$, which is part of the latent field, INLA methods are not directly applicable. Martino et al. [2011] note that this log-likelihood contribution is equal to the log-likelihood of a Poisson regression model with k Poisson-distributed data points, which can be used to cast the model into a LGM framework.

In this Poisson model, there is one data point with mean $(t - s_k)\exp(\eta_k)$ that is 1 or 0 depending on whether the observation is censored, while $k - 1$ data points

are observed to be 0 with mean $(s_{j+1} - s_j)\exp(\eta_j)$. The dataset is augmented so that k Poisson-distributed data points represent each data point (t, δ). This model is a LGM, making it possible to apply the INLA methodology.

4.3 Model specification

We assume a Cox PH model for the survival time after MI. This analysis only applies to the 4,248 participants who have had an MI. The observed variables $\boldsymbol{y}$ is the length of follow up time t, as well as an indicator variable δ stating whether the participant has died (value 1) or if the time has been censored (value 0).

The dataset includes the possible effects of age, sex, and season. Age is a discrete variable giving the age at MI in years. Sex is an indicator variable that is 0 for women and 1 for men, while season is an indicator variable that is 0 for winter (months November-January) and 1 for not-winter (months February-October). We fit the models with the following predictors for patient i to check whether the proportionality assumption holds, and find the model with the best fit:

$$
\begin{aligned}
&1) \quad \eta_{ik} = b_k + \beta_0 + f^{(\text{age})}(\text{age}_i) + \beta_{\text{sex}}\text{sex}_i + \beta_{\text{season}}\text{season}_i \\
&2) \quad \eta_{ik} = b_k + \beta_0 + f^{(\text{age})}(\text{age}_i) + \beta_{\text{sex}}\text{sex}_i \\
&3) \quad \eta_{ijk} = b_k^j + \beta_0 + f^{(\text{age})}(\text{age}_i) + \beta_{\text{sex}}\text{sex}_i \\
&4) \quad \eta_{ijk} = b_k^j + \beta_0 + f^{(\text{age})}(\text{age}_i) + \beta_{\text{season}}\text{season}_i \\
&5) \quad \eta_{ijk} = b_k^j + \beta_0 + f^{(\text{age})}(\text{age}_i)
\end{aligned}
$$

Here, $i = 1, \ldots, n$, $k = 1, \ldots, 33$ and $j = 1, 2$. In model 3, season is stratified. Sex is the stratum variable in models 4 and 5. The baseline hazard $h_0(t)$ is modelled using a RW1 model, as given in equation (3.16), with precision parameter τ_b. The precision parameter is assigned a PC prior. The timeline is partitioned in 33 parts, one for each year.

An age effect is included in all model configurations, and modelled using a RW2 prior, as given in equation (3.16). The prior has precision parameter τ, and is assigned the PC prior in equation (3.18). The model has been scaled according to Sørbye and Rue [2014].

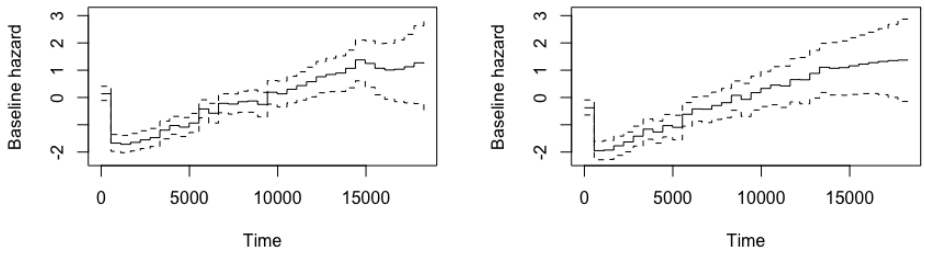
4.4 Results

The different models were compared using the DIC and the mean LS, given in equations (3.19) and (3.20). The results are shown in table 4.1. To analyse effects in a Cox PH model, the proportionality assumption must hold for each effect. To check that this assumption holds, a model where the variable is stratified can be compared to a corresponding simple model [Martino et al., 2011].

Table 4.1: DIC and mean LS values for Cox PH models with different effects included.

Model	Sex	Season	DIC	$\overline{\text{LS}}$
1	Linear	Linear	29262.61	0.56158
2	Linear	Not included	29260.60	0.56154
3	Linear	Stratified	29256.65	0.56147
4	Stratified	Linear	29255.65	0.56145
5	Stratified	Not included	29253.62	0.56141

The model with the lowest DIC and mean LS is model 5, where men and women are modelled as two separate sub-populations, and a seasonal effect is not included. The estimated baseline hazards for men and women are shown in figure 4.1. The curves are similar, but as the credible intervals do not include both baseline hazards, it is clear that they are significantly different. In addition, the largest difference between the two baseline hazards is when the time is low and the number of active participants is highest, indicating that the difference between men and women is not caused by an uncertain estimate due to a small number of participants. We conclude that the proportionality assumption does not hold for the sex effect, and inference is performed with model 5.



(a) **(b)**

Figure 4.1: Estimated baseline hazards for men (a) and women (b) in model 5, with 0.025 and 0.975 quantiles.

Table 4.2 shows the summary statistics for the intercept and the hyperparameters for model 5. There was a small positive effect for age (the hazard is greater with older age) and the baseline hazard is higher for men. The baseline hazard for men and women is shown in figure 4.1. For both men and women, the hazard rate is higher in the time right after a MI, before it drops. Then, it rises with time, due to the age effect. The age effect is assumed to be joint for both men and women, and is shown in figure 4.2. It shows a higher mortality hazard rate with older age, indicating a higher risk of death.

Table 4.2: The mean, standard deviation (SD), 2.5 %, 50 % and 97.5 % quantiles, and the mode of the effects in model 5.

The best fitted model, with sex as a stratum variable, and without a seasonal effect.

Linear effects	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Intercept	-7.395	0.126	-7.658	-7.39	-7.163	-7.38
Hyperparameters	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Precision for $h_0(t)$	0.966	0.228	0.587	0.943	1.48	0.901
Precision for age	30.218	22.999	5.359	24.289	90.87	14.274

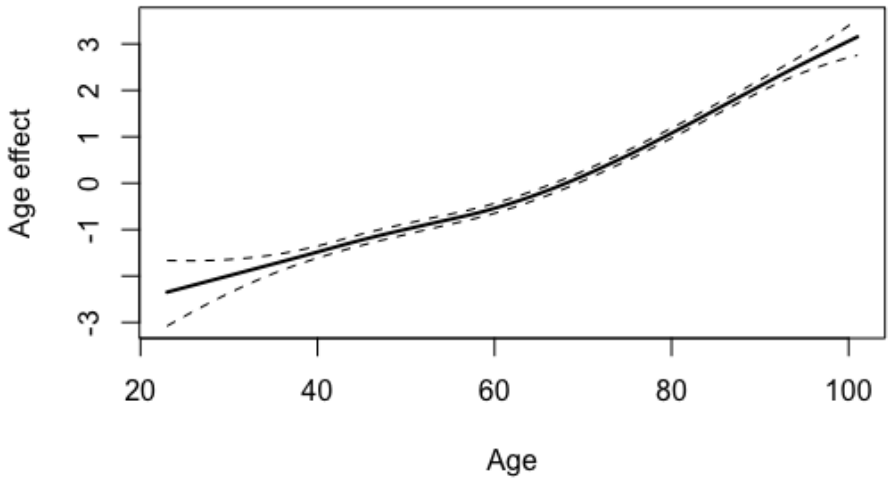


Figure 4.2: The posterior mean of the age effect, with 0.025 and 0.975 quantiles.

4.5 Discussion

To test the proportionality assumption, both stratified and simple models were fitted and compared using the DIC and mean LS, as well as a comparison of the estimated mortality hazard rates. The estimated mortality hazard rates for men and women were similar, but the credible intervals did not cover both hazard rates. As the estimated mortality hazard rates for men and women in the stratified model differed, and the stratified model had a lower DIC and mean LS, sex was modelled as a stratum variable.

The models without the seasonal effect had a lower DIC and mean LS than the models including the seasonal effect. Therefore, the seasonal effect was excluded from further analysis, as it did not contribute significantly to the analysis of this dataset. This indicates that the season does not affect the mortality hazard rate after a MI. A potential trend of weather or season in specific groups, such as age groups above 65 years or for men, has not been studied due to the scope of this thesis.

There is an increasing effect with age, where older participants have a higher mortality hazard rate. The credibility bands are wider at the youngest and oldest ages due to the small number of participants. The age trend has a higher slope after approximately 65 years of age, indicating that the risk changes quicker with age in older age groups.

The mortality hazard rate is high right after a MI, before dropping markedly shortly after. This drop indicates that the increased risk of death after an MI does not have a lasting effect. After this drop, the mortality hazard rate rises gradually with time, as participants age. An eventual change in the mortality hazard rate after a MI has not been studied due to the scope of this thesis.

5

The Bayesian age-period-cohort model

In chapter 2, an introductory analysis of the incidence rate of MI showed that the number of MIs increase with time as the mean age and the number of participants in the study increases. More advanced methods are needed to study the rate, to adjust for the increasing age of participants. One such method is the BAPC model. This model describes vital rates using age (the age at diagnosis), period (the date at diagnosis) and cohort (the date of birth). In this chapter, the BAPC model is presented, along with its application. It is used to study the change in the rate of MIs. Both an age-period (AP) model and an APC model have been fitted. The results are presented and discussed.

5.1 The BAPC model

Riebler and Held [2017] define the univariate APC model by

$$\begin{aligned} y_{ij} \mid \eta_{ij} &\sim \text{Poisson}(n_{ij} \exp(\eta_{ij})) \\ \eta_{ij} &= \alpha + f^{(\text{A})}(\text{A}_i) + f^{(\text{P})}(\text{P}_j) + f^{(\text{C})}(\text{C}_k) + f^{(\text{OD})}(\text{OD}_{ij}) \end{aligned} \tag{5.1}$$

The model includes the age effect A, the period effect P, the cohort effect C, as well as the effect OD_{ij} that accounts for possible overdispersion (OD). Here, $i = 1, \ldots, I$ is the number of age groups, $j = 1, \ldots, J$ is the number of periods, and $k = M \times (I - i) + j$ is the cohort index. M is the ratio between the length of the age groups and the length of the periods.

In a LGM framework, the observations are denoted $\boldsymbol{y}$, and the latent field is $\boldsymbol{x} = [\boldsymbol{\eta}, \alpha, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{OD}]$. The hyperparameters are the precision parameters $\boldsymbol{\theta} = [\tau_{\text{A}}, \tau_{\text{B}}, \tau_{\text{C}}, \tau_{\text{OD}}]$.

5.1.1 Identifiability problem

As age, period, and cohort are linearly dependent on each other, there is a known identifiability problem in APC analysis. There are infinitely many combinations giving the same estimated rate, and additional constraints are required to ensure identifiability [Riebler and Held, 2017].

Some common constraints are to apply sum-to-zero constraints $\sum_i \text{A}_i = \sum_j \text{P}_j = \sum_k \text{C}_k = 0$, allowing the identification of the intercept [Riebler and Held, 2017]. However, it is still not possible to identify the effects of age, period, and cohort. It is possible to identify second differences for the age, period, and cohort effects. For the age effect, these second differences are given as

$$(A_{i+1} - A_i) - (A_i - A_{i-1}) = A_{i+1} - 2A_i + A_{i-1} \tag{5.2}$$

The second differences for period and cohort effects are found analogously. The second differences are identifiable and represent the curvature of the APC effects, allowing the identification of trend changes [Riebler and Held, 2017]. The second differences can also represent ratios of relative risks. The relative risk for two age groups describes the change in risk from one age group to another, and the second differences represent the rate of change for the relative risk [Riebler and Held, 2017].

Rosenberg and Anderson [2011] argue that this identifiability problem is not unique to the APC model, and that the analysis of any cohort study will have a similar identifiability problem or "uncertainty principle". They note that the APC model can be partitioned into linear and non-linear components in several useful ways, such as the AP form or the age-cohort (AC) form. This will allow the identification of several relevant parameters, such as longitudinal or cross-sectional age trends.

5.2 Model specification

Due to the number of participants, the MI incidences before 1982 and at age above 79 years have not been included in this analysis. Due to the gender differences shown in chapter 2, only men have been included in the model. This gives a total of 19,627 participants, of which 2,385 have experienced a MI.

The resulting analysis is performed on a dataset with 8 five-year age groups (40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, and 75-79 years) over a period of 33 years from 1982 to 2014. The cohort is calculated by $k = 5 \cdot (8 - i) + j$, as the ratio between the length of the age groups and periods is 5.

The data is represented as in table 5.1, which shows the MI and participant count for the years 1995-1999. The structure of the data is analogous for the other years in the study.

Table 5.1: The data set is represented in a matrix with the relevant counts in each year sorted by age group.

MI counts								
	40-44	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**
1995	1	8	12	9	13	12	9	11
1996	2	9	12	11	10	8	7	13
1997	2	9	15	16	12	16	11	12
1998	2	11	7	20	13	18	11	9
1999	3	8	13	16	12	20	18	12
Participant counts								
	40-44	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**
1995	2874	2728	2128	1361	1141	1091	576	350
1996	3059	2848	2245	1412	1162	1086	502	391
1997	3237	2921	2343	1497	1177	1105	421	413
1998	3324	3014	2466	1554	1181	1119	352	451
1999	2424	2700	2676	1960	1302	1115	1180	524

5.2.1 AP model

An AP model allows the identification of the net drift, the parameter $(P_L + C_L)$, where P_L is the linear effect of period and C_L is the linear effect of successive cohorts [Rosenberg and Anderson, 2011]. This is the joint effect of period and cohort. It also allows the identification of a cross-sectional age trend $(A_L - C_L)$,

where A_L is the linear age effect and C_L is the linear effect of successive cohorts.

The fitted model has predictor

$$\eta_{ij} = \alpha + f^{(\mathrm{A})}(\mathrm{A}_i) + f^{(\mathrm{P})}(\mathrm{P}_j) + f^{(\mathrm{OD})}(\mathrm{OD}_{ij})$$

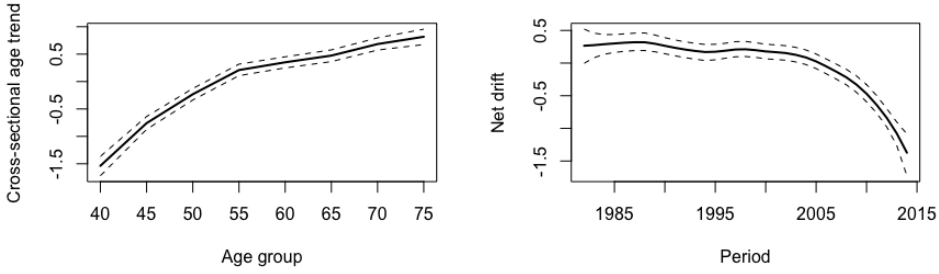
The age effect A and the period effect P are fitted with RW2 models, as given in equation (3.16). The OD effect is fitted with an independent random noise model, as given in equation (3.17). All hyperparameters are assigned PC priors.

5.2.2 APC model

An APC model is fitted with the predictor given in equation (5.1). The age, period, and cohort parameters are assigned RW2 priors, while the OD parameter is assigned an independent random noise prior. All hyperparameters are assigned PC priors.

5.3 Results

Figure 5.1 shows the effects in the AP model, the cross-sectional age trend and the net drift. The cross-sectional age trend shows an increasing MI rate with older age, while the net drift shows a decreasing rate with period.



(a) **(b)**

Figure 5.1: The effects in the AP model: the cross-sectional age trend (a) and the net drift (b), with 0.025 and 0.975 quantiles.

The summary statistics of the intercept and the precision parameters for the AP and APC model are given in Table 5.2.

Table 5.2: The mean, standard deviation (SD), 2.5 %, 50 % and 97.5 % quantiles, and the mode of the effects in the AP and APC model.

AP model						
Linear effects	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Intercept	-5.128	0.029	-5.186	-5.128	-5.073	-5.127
Hyperparameters	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Precision for age	34.84	29.30	5.92	26.75	112.10	15.19
Precision for period	14.45	14.41	1.66	10.25	52.64	4.52
Precision for OD	28.83	12.17	13.48	26.03	59.94	21.56
APC model						
Linear effects	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Intercept	-5.227	0.055	-5.343	-5.224	-5.126	-5.219
Hyperparameters	**Mean**	**SD**	**0.025 Q**	**0.5 Q**	**0.975 Q**	**Mode**
Precision for age	35.43	29.24	5.92	27.47	112.94	15.48
Precision for period	14.76	15.27	1.55	10.28	55.12	4.27
Precision for cohort	38.32	49.31	3.50	23.53	164.53	9.18
Precision for OD	39.64	21.81	15.70	33.80	96.85	25.71

Figure 5.2 shows the second differences for the age, period, and cohort effects. The period and cohort effects have second differences close to 1 for all periods and cohorts, indicating that the relative risk has a constant rate of change. The second differences do not give info as to whether there is a steadily increasing or decreasing trend of the rate, or if the rate is constant across all periods and cohorts.

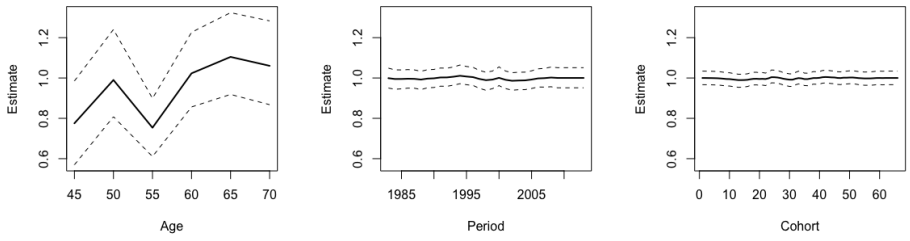


Figure 5.2: Mean, 2.5 % quantile and 97.5 % quantile of the identifiable second differences on exponential scale.

There is a change in the curvature of the age trend around 45-49 years and 55-59 years. For the other age groups, the second differences have values close to 1, and the ratio of the relative risks between these age groups and their

neighbours have a constant rate of change. As the second differences for age are below 1 for age groups 40-44 to 60-64, the ratio of the relative risks decreases between these age groups.

Table 5.3 shows the DIC and mean LS for an AP and an APC model. As the APC model has a lower DIC and mean LS, this model is selected for analysing the rate of MI. The identification problem of APC models does not affect rate estimates.

Figure 5.3 shows the projected age-standardised rates of MI per 1000 participants, while figure 5.4 shows the projected age-specific rates of MI per 1000 participants for age groups 40-44 to 75-79. The standardisation weights for the age-standardised rates were the number of MIs in each age group. The dots are the observations, while the outer edges of the fan shows the 0.025 and 0.975 quantiles. The bands of the fan show quantiles increasing by 10 % between these extremities.

Table 5.3: DIC and mean LS values for BAPC models with and without cohort.

Model	DIC	$\overline{\text{LS}}$
AP	1356.14	2.613
APC	1347.16	2.592

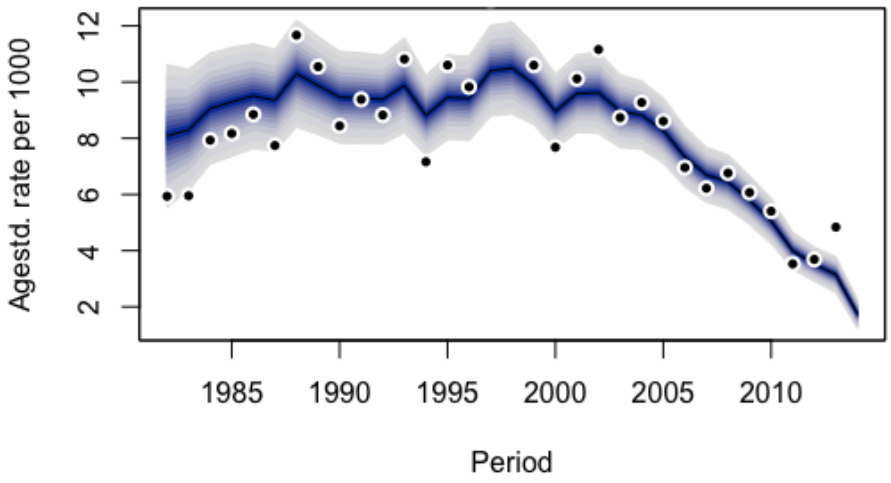


Figure 5.3: Age-standardised MI rates, with a fan showing the 0.025 and 0.975 quantiles, and quantiles in 10 % increments within this interval.

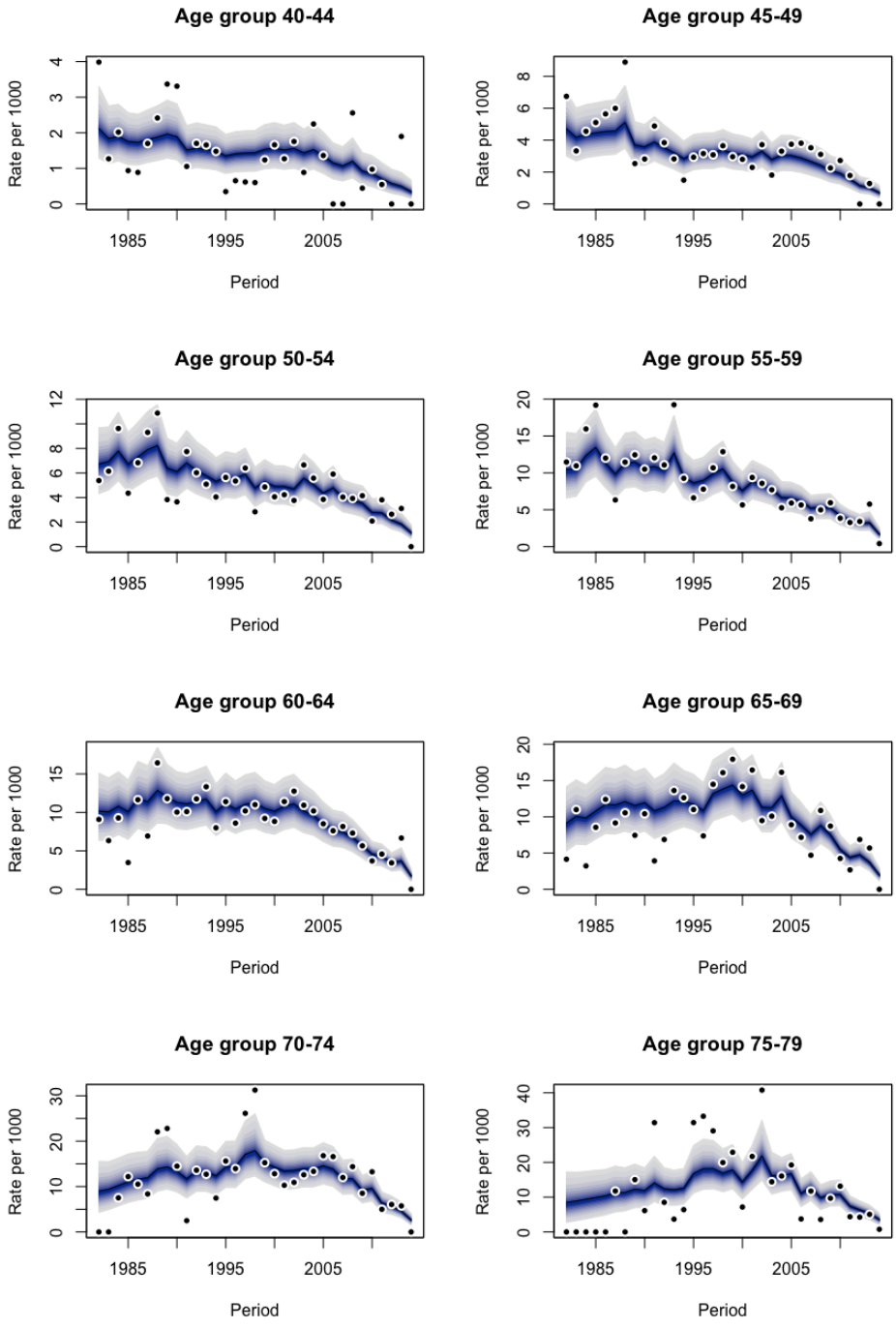


Figure 5.4: Age-specific MI rates for age groups 40-44 to 75-79, with a fan showing the 0.025 and 0.975 quantiles, and quantiles in 10 % increments within this interval.

5.4 Discussion

Although our dataset covers 39,870 participants, only 4,248 of these have experienced a MI. When studying the incidence rate of MI in the different age groups, this required splitting our dataset into several smaller groups, leaving some of them with few participants. This affects the rate estimation, as the estimated rate is more sensitive to outliers, and the credible intervals are wider in age groups with fewer participants. The age groups above 75-79 years have been excluded from the analysis, as the sample size was too small to give reliable estimates.

The AP model estimated an increasing cross-sectional age trend and decreasing net drift, as shown in figure 5.1. The slope of the cross-sectional age trend is flatter in higher age groups, while the rate of change of the net drift increases with time. This change in the slope of the rate in the net drift is not visible in the second differences, as the second differences for both period and cohort are approximately constant and equal to 1. However, the decrease in rate with period is consistent with the projected incidence rates of MI.

The second differences for age shown in figure 5.2 show a change in the ratios between the relative risks for age groups 45-49 and 55-59. This is consistent with the cross-sectional age trend, where the slope of the trend flattens slightly both at age group 45-49 and 55-59 in figure 5.1.

The age-standardised and age-specific rates are shown in figures 5.3 and 5.4. The age-standardised rates show that the overall rate of MI was approximately constant for much of the time of the study, before decreasing after year 2000. This is consistent with the trend seen for the net drift. These age-specific rates can be compared with the rates shown in figure 2.4, which merely divided the number of MIs by the number of participants. Here, the effects have been smoothed, and a trend is easier to identify.

The decreasing trend of MI during the time of the study is seen across all age groups. The estimated incidence rates of MI are higher in older age groups. This is consistent with the cross-sectional age trend. The four oldest age groups have a slight increasing trend in the earlier years of the study. This may be due to the low number of participants in these age groups in the early years of the study, leading to artificially low MI counts for these years.

Introductory analysis showed that the mean age at first MI in the Tromsø Study was approximately 65 years. This result is not seen when corrected for the number of participants, and the MI incidence rate rises strictly with age.

These results can be compared to the results presented by the Norwegian Institute of Public Health [2009]. Our dataset also showed that the number of first-time MIs have increased in the age groups over 65 years, but this trend is seen in all the age groups of our dataset. The Norwegian Institute of Public Health [2009] also reported a worrying MI trend in the age group 24-44 that can not be verified or falsified in our dataset, as we do not have a sufficient number of MI incidents at these ages to perform analysis.

6

Conclusion

6.1 Summary

The prevalence of cardiovascular diseases in Norway makes a study of the rate of incidences and mortality an interesting topic. In addition to established risk factors, there are several myths about MIs, such as snowfall leading to an increased risk of MI. The analyses were performed on a combined dataset from the Tromsø Study dataset, which contained information about sex, date of first MI and date of attendance in the Tromsø Study, and weather data from the Norwegian Meteorological Institute, which contained information about the temperature and snowfall at the date of MI.

Introductory analysis did not show significant effects for either temperature or snowfall. The results for temperature are consistent with previous results for MIs in northern climates [Hopstock, 2012, Mohammad et al., 2018, Barnett et al., 2005], while the results for snowfall are not consistent with the findings of [Auger et al., 2017]. This analysis was based on data of the weather in Tromsø on the day of MI. As there was not found an effect of temperature or snowfall on MI, season was determined by month (November-February were counted as winter months) in further analysis.

The topics chosen for further analysis was the mortality hazard rate after having an MI and the incidence rate of MI. These topics were studied using a Cox PH model and a BAPC model, respectively. Inference was performed using the INLA methodology.

A stratified Cox PH model without a seasonal effect with sex as the stratum variable was deemed to be the best fit to analyse the mortality hazard rate after MI in this dataset. The models were compared using the DIC and the mean LS. The results showed that the hazard rate was high in the time shortly after a MI, before it dropped and rose steadily with age.

The incidence rate of MI was analysed by applying an APC model. An AP model was also applied, to identify the cross-sectional age trend and the net drift. The projected rates from the APC model showed that the rate of MI is higher in older age groups, but that the incidence rate is decreasing with time. The second differences of the APC model showed some changes in the curvature of the age trend, indicating that the rate of change of the age trend decreases. This was consistent with the cross-sectional age trend found in the AP model.

The second differences of the cohort and period trends indicated a constant rate of change in the cohort and period trends. This is not consistent with the net drift shown by the AP model, which show a change in the curvature of the trend after approximately year 2005. Until this point, the net drift trend had been approximately constant, before it started decreasing. This decreasing trend is consistent with the trend seen in the projected rates.

The trends of the incidence and mortality hazard rates were consistent with the hypotheses presented in the introduction. The mortality hazard rate was higher in the days after an MI, before dropping and increasing with time as the participants age. The mortality hazard rate for women and men had approximately the same shape, but the rate was slightly lower for women, especially in the days immediately following the MI.

The incidence rate of MI was expected to decrease during the time of the study, and the rate was expected to be higher for men than women. The incidence rate was expected to be influenced by weather. Introductory analysis showed that the rate of MIs was different for men and women, that a higher proportion of men had an MI, and that there was no link between temperature or snowfall and MI. For further analysis, only the incidence rate of men were studied, using a BAPC model. The incidence rate decreased during the time of the study as expected.

In conclusion, the analyses in this thesis did not show a connection between temperature or snowfall and the MI incidence rate or the MI mortality hazard rate. The mortality hazard rate was higher in the time shortly after a MI before dropping and increasing with age. The incidence rate of MI was higher for older age groups, and decreased in the later years of the study.

6.2 Future research

There are several possible areas for further research into the prevalence of MI in Tromsø. This thesis did not account for other known risk factors for cardiovascular disease, and the number of questions analysed were limited both by the data available and the scope of this thesis.

A further study of the snow shovelling hypothesis should be connected to more accurate data on which participants have shovelled snow. The analysis should also include MIs occurring in the days following a snowfall. It could also be interesting to differ between isolated days with snowfall, and several following days of snowfall.

Our dataset only included data for first-ever MIs. If patients who have already had an MI are more at risk due to colder temperatures or higher snowfall, this would not be visible in this dataset. The dataset also included a limited number of MI incidences in younger and older age groups, making the analysis for these age groups unreliable. A study of the incidence rates in these age groups requires more data.

Several questions were also left unstudied due to the scope of the thesis. A potential change in the mortality hazard rate with time was not studied. As the seasonal variation was not found to be statistically significant, it was excluded from further analysis. Hopstock [2012] found a seasonal variation, but only for those above 65 years. A potential seasonal variation in the mortality hazard rate for specific age groups has not been studied.

In the introductory analysis, the mean age at MI for the entire dataset was 65 years, while the mean age at MI increased with decade. This was deemed to be caused by the ageing population of participants. A potential change in the mean age at MI has not been studied.

As the introductory analysis showed that temperature and snowfall did not have an effect on the incidences of MI, it was not included in the APC analysis of the incidence rate. A further APC analysis could compare the incidence rates for different seasons to study potential effect of season.

A

Appendix

A.1 Cox PH model

```
library(INLA)

inla.setOption(scale.model.default = TRUE)
inla.setOption("enable.inla.argument.weights", TRUE)
hyper.pc = list(prec = list(prior = "pc.prec", param
    = c(1,0.01)))

#Model 1
formula.1 = inla.surv(time,cens) ~ f(age.mi,
   model="rw2", hyper=hyper.pc) + sex.mi + season.mi
result.1 = inla(formula.1, family="coxph", data =
   cox.data, control.hazard = list(model="rw1",
   hyper=hyper.pc, n.intervals=33), control.compute
   = list(dic=TRUE, cpo=TRUE))

#Model 2
formula.2 = inla.surv(time,cens) ~ f(age.mi,
   model="rw2", hyper=hyper.pc) + sex.mi
result.2 = inla(formula.2, family="coxph", data =
   cox.data, control.hazard = list(model="rw1",
   hyper=hyper.pc, n.intervals=33), control.compute
   = list(dic=TRUE, cpo=TRUE))
```

```
#Model 3
formula.3 = inla.surv(time,cens) ~ f(age.mi,
   model="rw2", hyper=hyper.pc) + sex.mi.mean
result.3 = inla(formula.3, family = "coxph", data =
   cox.data, control.hazard = list(model = "rw1", n.
   intervals = 33, strata.name = "season.mi"),
   control.compute = list(dic=TRUE, cpo=TRUE))

#Model 4
formula.4 = inla.surv(time,cens) ~ f(age.mi,
   model="rw2", hyper=hyper.pc) + season.mi
result.4 = inla(formula.4, family="coxph", data =
   cox.data, control.hazard = list(model = "rw1", n.
   intervals = 33, strata.name = "sex.mi"),
   control.compute = list(dic=TRUE, cpo=TRUE))

#Model 5
formula.5 = inla.surv(time,cens) ~ f(age.mi,
   model="rw2", hyper=hyper.pc)
result.5 = inla(formula.5, family="coxph", data =
   cox.data, control.hazard = list(model = "rw1", n.
   intervals=33, strata.name="sex.mi"), control.
   compute = list(dic=TRUE, cpo=TRUE))

#Recompute CPO values which violate assumptions
result.cox1 = inla.cpo(result.cox1)
result.cox2 = inla.cpo(result.cox2)
result.cox3 = inla.cpo(result.cox3)
result.cox4 = inla.cpo(result.cox4)
result.cox5 = inla.cpo(result.cox5)

#Calculate the mean LS
-mean(log(result.cox1$cpo$cpo),na.rm=T)
-mean(log(result.cox2$cpo$cpo),na.rm=T)
-mean(log(result.cox3$cpo$cpo),na.rm=T)
-mean(log(result.cox4$cpo$cpo),na.rm=T)
-mean(log(result.cox5$cpo$cpo),na.rm=T)

#Generate figure 4.1
eval.point = result.cox5$.arg$data$baseline.hazard.
   values
baseline.hazard = matrix(c(result.cox5$summary.
   random$baseline.hazard$mean,result.cox5$summary.
```

```
   random$baseline.hazard$`0.025quant`,result.cox5$
   summary.random$baseline.hazard$`0.975quant`) +
   exp(result.cox5$summary.fixed[1,1]), nrow =
   length(eval.point))

matplot(eval.point,baseline.hazard[,c(1,3,5)],type=
   type,col=1,lty=c(1,2,2),ylim=range(basehaz),xlab=
   "Time",ylab="Baseline hazard")
matplot(eval.point,baseline.hazard[,c(2,4,6)],type=
   type,col=1,lty=c(1,2,2),ylim=range(basehaz),xlab=
   "Time",ylab="Baseline hazard")

#Generate figure 4.2
matplot(result.cox5$summary.random$age.mi$ID,
cbind(result.cox5$summary.random$age.mi$mean,result.
   cox5$summary.random$age.mi$`0.025quant`,result.
   cox5$summary.random$age.mi$`0.975quant`),
type="l",lwd=c(2,1,1),lty=c(1,2,2),col=1,xlab="Age",
   ylab="Age effect")
```

A.2 BAPC model

```
library(INLA)
library(BAPC)

#AP Model
result.ap = BAPC(mi.men.bapc, predict=list(npredict
   =0), model = list(age=list(model="rw2",prior="pc.
   prec",param=c(1,0.01),scale.model=T), period=list
   (include=T,model="rw2",prior="pc.prec",param=c
   (1,0.01),scale.model=T), cohort=list(include=F)))
inla.ap = inlares(result.ap)

#APC Model
result.apc = BAPC(mi.men.bapc, predict=list(npredict
   =0), model = list(age=list(model="rw2",prior="pc.
   prec",param=c(1,0.01),scale.model=T), period=list
   (include=T,model="rw2",prior="pc.prec",param=c
   (1,0.01),scale.model=T), cohort=list(include=T,
   model="rw2",prior="pc.prec",param=c(1,0.01),scale
   .model=T)),secondDiff = T,stdweight=colSums(epi(
   mi.men.bapc)))
```

```
10 inla.apc = inlares(result.apc)
11
12 #Recompute CPO values which violate assumptions
13 result.ap = inla.cpo(result.ap)
14 result.apc = inla.cpo(result.apc)
15
16 #Calculate the mean LS
17 -mean(log(result.ap$cpo$cpo),na.rm=T)
18 -mean(log(result.apc$cpo$cpo),na.rm=T)
19
20 #Generate figure 5.1
21 matplot(seq(40,75,5),cbind(inla.ap$summary.random$i$
      mean,inla.ap$summary.random$i$`0.025quant`,inla.
      ap$summary.random$i$`0.975quant`),type="l",pch
      =19,col=1,lty=c(1,2,2),lwd=c(2,1,1),xlab="Age",
      ylab="Effect")
22 matplot(rep(1982:2014),cbind(inla.ap$summary.random$
      j$mean,inla.ap$summary.random$j$`0.025quant`,inla
      .ap$summary.random$j$`0.975quant`),type="l",pch
      =19,col=1,lty=c(1,2,2),lwd=c(2,1,1),xlab="Period"
      ,ylab="Effect")
23
24 #Extract the second differences
25 sec.diff.age=summarySecDiff(result.apc.all,variable=
      "age")
26 sec.diff.per=summarySecDiff(result.apc.all,variable=
      "period")
27 sec.diff.coh=summarySecDiff(result.apc.all,variable=
      "cohort")
28
29 #Generate figure 5.2
30 matplot(seq(45,70,5),sec.diff.age[,c(1,3,5)],type="l
      ",col=1,lwd=c(2,1,1),lty=c(1,2,2),xlab="Age",ylab
      ="Estimate")
31 matplot(rep(1983:2013),sec.diff.per[,c(1,3,5)],type=
      "l",col=1,lwd=c(2,1,1),lty=c(1,2,2),xlab="Period"
      ,ylab="Estimate",ylim=range(sec.diff.age[,c
      (1,3,5)]))
32 matplot(sec.diff.coh[,c(1,3,5)],type="l",col=1,lwd=c
      (2,1,1),lty=c(1,2,2),xlab="Cohort",ylab="Estimate
      ",ylim=range(sec.diff.age[,c(1,3,5)]))
33
34
```

```
35 #Generate figure 5.3
36 plotBAPC(result.apc,type="ageStdRate",scale=10^3)
37
38 #Generate figure 5.4
39 plotBAPC(result.apc,type="ageSpecRate",scale=10^3)
```

Bibliography

N. Auger, B. J. Potter, A. Smargiassi, M. Bilodeau-Bertrand, C. Paris, and T. Kosatsky. Association between quantity and duration of snowfall and risk of myocardial infarction. *Canadian Medical Association Journal*, 189(6): E235–E242, 2017.

A. G. Barnett, A. J. Dobson, P. Mcelduff, V. Salomaa, K. Kuulasmaa, and S. Sans. Cold periods and coronary events: an analysis of populations worldwide. *Journal of Epidemiology and Community Health*, 59(7), 2005.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer, Berlin, 2nd edition, 2001.

T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

I. J. Good. Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952.

L. Held, B. Schrödle, and H. Rue. Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, pages 91–110. Physica-Verlag HD, Heidelberg, 2010.

L. A. Hopstock. *Seasonal variation in incidence of acute myocardial infarction and cardiovascular disease risk factors in a subarctic population : the Tromsø Study*. PhD thesis, University of Tromsø, Tromsø, 2012.

B. K. Jacobsen, A. E. Eggen, E. B. Mathiesen, T. Wilsgaard, and I. Njølstad. Cohort profile: The Tromsø Study. *International Journal of Epidemiology*, 41 (4):961–967, 2012.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

S. Martino, R. Akerkar, and H. Rue. Approximate Bayesian Inference for Survival Models. *Scandinavian Journal of Statistics*, 38(3):514–528, 2011.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67(C): 68–83, 2013.

M. A. Mohammad, S. Koul, R. Rylance, O. Fröbert, J. Alfredsson, A. Sahlén, N. Witt, T. Jernberg, J. Muller, and D. Erlinge. Association of Weather with Day-to-Day Incidence of Myocardial Infarction: A SWEDEHEART Nationwide Observational Study. *JAMA Cardiology*, 3(11):1081–1089, 11 2018.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Norwegian Institute of Public Health. Cardiovascular disease in Norway, 2009. URL `https://www.fhi.no/en/op/hin/health-disease/cardiovascular-disease-in-norway---/`.

Norwegian Ministry of Education and Research. Læreplan i matematikk for samfunnsfag - programfag i utdanningsprogram for studiespesialisering, 2006. URL `http://data.udir.no/kl06/MAT4-01.pdf`.

Norwegian Ministry of Education and Research. Læreplan i matematikk fellesfag, 2013. URL `http://data.udir.no/kl06/MAT1-04.pdf`.

Norwegian Ministry of Education and Research. Overordnet del – verdier og prinsipper for grunnopplæringen, 2017. URL `https://www.regjeringen.no/contentassets/37f2f7e1850046a0a3f676fd45851384/overordnet-del---verdier-og-prinsipper-for-grunnopplaringen.pdf`.

L. I. Pettit. The Conditional Predictive Ordinate for the Normal Distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184, 1990.

M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

A. Riebler and L. Held. Projecting the future burden of cancer: Bayesian age–period–cohort analysis with integrated nested Laplace approximations.

Biometrical Journal, 59(3):531–549, 2017.

P. S. Rosenberg and W. F. Anderson. Age-period-cohort models in cancer surveillance research: ready for prime time? *Cancer epidemiology, biomarkers & prevention*, 20(7):1263–1268, 2011.

H. Rue and L. Held. *Gaussian Markov random fields : theory and applications*, volume 104 of *Monographs on statistics and applied probability*. Chapman & Hall/CRC, Boca Raton, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1–28, 2017.

S. H. Sørbye and H. Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8(C):39–51, 2014.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81 (393):82–86, 1986.

X. Wang, Y. Yue Ryan, and J. J. Faraway. *Bayesian Regression Modeling with INLA*. Chapman & Hall/CRC. CRC Press, Boca Raton, 1st edition, 2018.

S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos, and L. Lisheng. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*, 364(9438):937–952, 2004.