Mathematical Statistics
Stockholm University

Estimating the heritability of survival time after acute myocardial infarction using population-based national Swedish health registries

Sara Ekberg

Examensarbete 2011:1

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2011:1**,
http://www.math.su.se/matstat

Estimating the heritability of survival time after acute myocardial infarction using population-based national Swedish health registries

Sara Ekberg[*]

Februari 2011

Abstract

Acute myocardial infarction (AMI) is the most common cause of death in Sweden. The aim of this study was to assess the heritability of survival time after acute myocardial infarction in full siblings. The study was based on reported incidents of Acute Myocardial Infarction in Sweden between 1987 and 2006 to either the National Patient Register or the Cause of Death register. We used the Multigenerational register to identify full sibling pairs where both had suffered from AMI. In this study we are focusing on the nonimmediate deaths (i.e. patients surviving the first day after AMI). Three different outcomes were studied: overall mortality, cause-specific death and repeated AMI. For each different outcome a Cox proportional hazards model was fit to the whole population (second sibling in each sib-pair to suffer from AMI excluded), taken into account possible confounders i.e. age, sex, calendar year and county. These models served as adjusted baseline for average survival after an AMI event, from which we computed residuals for all the members of sib pairs that we were interested in. These residuals served as a quantitative, adjusted measure of prognosis, i.e. better or worse than expected for the given combination of age, sex etc. We then fitted a Cox proportional hazards model based on the second sibling in each sib-pair to suffer from AMI, using the first sibling's prognosis as exposure. For the outcome overall mortality the results indicate that there is an association between full siblings survival time but for the other two outcomes there is no evidence of association between full siblings survival times. This result indicates that the co-morbidity that we see for the outcome overall mortality can be due to shared frailties rather than a direct consequence of the AMI event.

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:ekberg.sara@gmail.com. Supervisor: Ola Hössjer.

Acknowledgement

I would like to thank my supervisors from the department of Medical Epidemiology and Biostatistics at Karolinska Institutet: Dr. Anna Bennet for contributing with her knowledge in epidemiology and for her encouragement and Dr. Alexander Ploner for his many hours of stimulating discussion and inexhaustible effort in helping this project forward. I would also like to thank my supervisor from Stockholm University Professor Ola Hössjer for reading and giving me valuable feedback. I am also thankful to Christer Platzack who has read and commented upon an earlier version of my thesis.

Contents

List of Figures

List of Tables

1 Introduction

Myocardial infarction is the most common cause of death, and one of the most common causes of invalidity in Sweden[11]. The aim of this project is to assess the heritability of survival time after acute myocardial infarction. Addressing this question requires very large clinical material with information about siblings or parent-child relationships. In this project data from the National Patient register, the Cause of Death register and the Multigenerational register have been linked together. We identify full sib-pairs from the multi-generation registry where both individuals have suffered an acute myocardial infarction.

1.1 Acute myocardial infarction

Acute myocardial infarction or heart attack is a heart condition where blood cannot reach parts of the heart and the heart muscle becomes damaged.

Myocardial infarction is often a result of atherosclerosis, when plaque is buildup over time on the inside of the walls of the coronary arterys. Rapture of an atherosclerotic plaque in the wall of the artery can cause a blood clot to block the coronary artery for oxygen-rich blood. The restriction of blood supply (ischemia) damages the heart muscle. If the ischemia has lasted too long the muscle will not recover[8]

In 2006 more than 36 000 individuals suffered from an acute myocardial infarction in Sweden[11]. The most important risk factors are smoking, high levels of lipoproteins, psychosocial factors, central obesity, high blood pressure and adult onset diabetes[12]. Almost one third of the persons diagnosed with myocardial infarction die within 28 days[12].

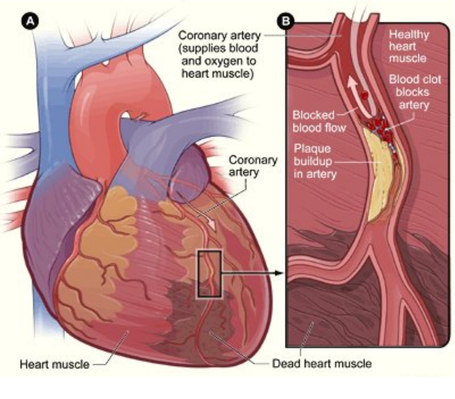


Figure 1: Illustration of Acute Myocardial Infarction (taken from U.S. Department of Health & Human Services, www.nhibi.se)

1.2 Background to this study

Studies show that differences in mortality with respect to gender vary with age. Younger women have worse prognosis than men[7][6]. Men are more likely to die before hospitalization [7][6]. During 1987 to 2006 the 28-day case fatality was reduced by almost two thirds in patients younger than 75 years old [1] This can be explained by an improved care program that for example involves faster surgical interventions and also by the introduction of new treatments such as intravenous beta-blockers, nitroglycerin infusion, aspirin and thrombolytics.

Several different factors have been suggested to influence the long-term prognosis after myocardial infarction in an adverse manner. For patients that survive at least 28 days, one of the most important predictor of recurrent myocardial infarction has been shown to be diabetes in both sexes. Other primary risk factors are job strain, central obesity in men, low level of A1 apolipoprotein and high-density lipoprotein cholesterol in women [4].

It is currently unknown if survival time/prognosis after acute myocardial infarction has a genetic component (i.e. is inherited). A positive family history of coronary heart disease is considered an independent risk factor for developing coronary heart disease. Patients with positive family history of coronary heart disease develop their first acute myocardial infarction earlier in comparison to patients without such history and have a better prognosis which is mostly explained by their lower age [2].

1.3 Report structure

In the present study we link together data from the National Patient Register, the Cause of Death Register and the Multigenerational register. The linkage of data sources are presented in section 2 and the resulting cohort characteristics are shown in Table 1 and 2. Section 3 begins with a short theoretical introduction to the Cox proportional hazards model and the assumption of proportional hazards. The section is closed by a scheme of how we use this model to address the question of heritable prognosis. In Section 4 we present the results of the analysis outlined in Section 3 for three different outcomes: overall mortality, cause-specific death and repeated infarction. Interpretations of the results are discussed in Section 5. We also discuss potential problems with the model used and possible future research as a continuation to this study.

2 Linkage of data sources

Included in the study cohort are persons with acute myocardial infarction reported either to the National Patient Register or the Cause of Death Register.

The National Patient Register contains information about all persons admitted to any public hospital in Sweden. The registration is nationwide since 1987 but the register has been in existence since 1964[14]. The Cause of Death Register contains information about all persons with residence in Sweden at the time of their death since 1961 and is updated anually [14]. Persons with their first reported acute myocardial infarction before the year of 1987 were excluded from the study because of the incompleteness of the National Patient Register.

To identify the persons with acute myocardial infarction we used the ICD-codes (International statistical Classification of Diseases and related health problems) which are handed out by WHO and are the international standard for classifying diseases. The ICD-codes have been updated several times, ICD-8 was used during the time period 1969-1986, ICD-9 during the period 1987-1996 and ICD-10 from 1997 is still used. We used ICD-8 and ICD-9 codes 410 and 411 and ICD-10 code I21 to identify persons with acute myocardial infarction from the National Patient register and the Cause of Death register. Each person enters the study on the date of their first myocardial infarction event and leaves the study at death, end of study or emigration. If a person has repeated events within the first 28 days from the first event this count as the same event. The first repeated event after a time period of 28 days is recorded as the second event. Persons with the first reported acute myocardial infarction before the age of 40 were excluded from the study.

The information about death date and cause of death are recorded from the Cause of Death register. The three different outcome that we will look at are overall mortality, cause-specific death and repeated infarction. We define the three different outcomes as:

- Overall mortality

  The ith individual is uncensored if reported dead to the Cause of Death register during the time of the study and censored otherwise.

- Cause-specific death

  The ith individual is uncensored if reported dead to the Cause of Death register with acute myocardial infarction as underlying cause of death (ICD-9: 410,411, ICD-10: I21) and is censored otherwise.

- Repeated infarction

  The *i*th individual is uncensored if he or she has a reported second myocardial infarction reported to either the National Patient register or the Cause of death register and is censored otherwise.

To identify the siblings we used the Multigenerational Register. The multigenerational register contains information about all persons with residence in Sweden since 1961 which are born 1932 or later. These persons are called index persons. The register contains information about connections between the index persons and their biological parents. The first version of the register was created in the year 2000. The register is updated anually. If more than two siblings in the same family have reported acute myocardial infarction either to the National Patient Register or the Cause of Death register we use the two oldest for consistency.

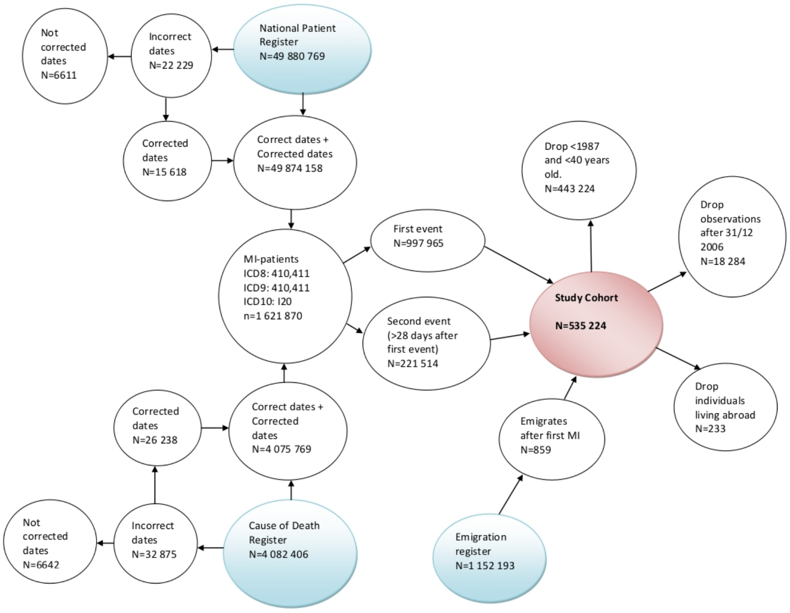

Not corrected dates N=6611
Incorrect dates N=22 229
National Patient Register N=49 880 769
Corrected dates N=15 618
Correct dates + Corrected dates N=49 874 158
Drop <1987 and <40 years old. N=443 224
Drop observations after 31/12 2006 N=18 284
First event N=997 965
MI-patients ICD8: 410,411 ICD9: 410,411 ICD10: I20 n=1 621 870
Study Cohort N=535 224
Second event (>28 days after first event) N=221 514
Drop individuals living abroad N=233
Corrected dates N=26 238
Correct dates + Corrected dates N=4 075 769
Emigrates after first MI N=859
Not corrected dates N=6642
Incorrect dates N=32 875
Cause of Death Register N=4 082 406
Emigration register N=1 152 193

Figure 2: Flowchart describing the data linkage

2.1 Missing or incomplete data

The National Patient Register contains over 49 million observations and contains information about both admission and discharge date. For consistency we use the admission date as the time of event. Less than 0.05 per mille of the observations (22 229) have missing values in either admission- or discharges dates. When admission date is incomplete but we have a complete discharge date we use the discharge date as the time of event. If month is missing from the admission date but we have it in the discharge date then we use the month from the discharge date. If the day is missing we use 15 as a proxy for the true date. After this operations there still remains 6611 observations with missing dates. We assume that these values are Missing Completely at Random (MCAR, see Little and Rubin,2002) [5], that is, if a value is missing or not is assumed to be independent of what we are studying and excluding the observations does not influence the results.

In the Cause of Death Register about 0.8 % has incomplete or missing date of death. If only the day was missing we used 15 as proxy for the true date. We identified the persons in the Cause of Death Register who died from a myocardial infarction and linked the two registers together. Since the same person can have several events we identify the first event for each person and consider this event as the entry date for that person.

2.2 Quality

The Swedish National Board of Health and Welfare (Socialstyrelsen) has presented a validation of the diagnosis codes in the National Patient Register for the years 1987 and 1995 with the purpose to detect possible bias in the reported myocardial infarction diagnosis due to differences in sex, age and geographic areas, hospitals and between specialized intensive care units and ordinary care units. According to the study the diagnosis for acute myocardial infarction is independent of age and sex, but there were some differences in the conformably of the definitions of the diagnosis codes and the clinical diagnosis in different counties and between specialized and non-specialized care units [13].

In the year 2001 the diagnosis acute myocardial infarction got a wider definition and as a consequence the number of reported acute myocardial infarctions increased [13]. Another source of error could be that the proportion of dissection of elderly non hospitalized individuals is low and the determination of the cause of death can therefore be inaccurate[11].

The Multigenerational Register has complete coverage since 1968 and good but yet not complete coverage since 1961.

2.3 Explanatory Variables

Information about age, sex, year and county is recorded at the entry day. Sweden is divided into 24 different counties and 290 municipalities. We use

SCB's (Statistics Sweden) classification of the municipalities into six different homogenous regions (H-regions). The regions are homogenous according to the catchment area. The scale ranges from H1 (big city) to H6 (sparsely populated area)[10]. The caracteristics of the full cohort is presented in Table 1 and the caracteristics of the sibling cohort is presented in Table 2.

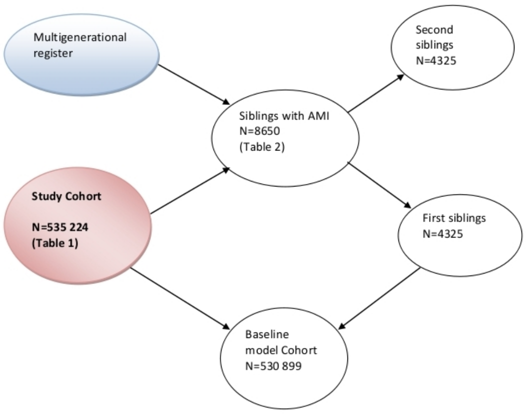

Multigenerational register
Second siblings N=4325
Siblings with AMI N=8650 (Table 2)
Study Cohort
N=535 224 (Table 1)
First siblings N=4325
Baseline model Cohort N=530 899

Figure 3: Flowchart describing the data linkage, continuation from Figure 1

Variable	Total	Crude death (%)	Cause-spec death(%)	Repeated infarction(%)
Age				
40-44	6558	1676 (25.6)	1120 (17.1)	1353 (20.6)
45-49	13787	3684 (26.7)	2451 (17.8)	2892 (21.0)
50-54	23085	6841 (29.6)	4440 (19.2)	4521 (19.6)
55-59	33216	12102 (36.4)	7625 (23.0)	6702 (20.2)
60-64	44997	21262 (47.2)	13096 (29.1)	9995 (22.2)
65-69	61256	36672 (59.9)	22006 (35.9)	14475 (23.6)
70-74	79493	56099 (70.6)	33992 (42.8)	18738 (23.6)
75-79	93241	74034 (79.4)	46039 (49.4)	20675 (22.1)
80-84	89267	76264 (85.4)	49199 (55.11)	17432 (19.5)
85-89	60701	54581 (89.9)	36872 (60.7)	9480 (15.6)
90-	30623	28445 (92.9)	20768 (67.8)	3300 (10.8)
Year				
87-91	148590	128636 (86.6)	80815 (54.4)	43090 (29.0)
92-96	146240	109761 (75.0)	68019 (46.5)	34732 (23.8)
97-01	121433	78541 (64.7)	50462 (41.6)	19914 (16.4)
02-06	119961	54722(45.6)	38312 (31.9)	11827 (9.9)
Sex				
male	318603	210919 (66.2)	135301 (42.5)	67555 (21.2)
female	217621	160741(73.9)	102307 (47.0)	42008 (19.3)
Region				
H1	76347	51945 (68.0)	33672 (44.1)	14571 (19.1)
H2	68651	46657(68.0)	28622 (41.7)	13913 (20.3)
H3	193600	132631 (68.5)	83937 (43.4)	40499 (20.9)
H4	113415	80518 (71.0)	52051 (45.9)	23510 (20.7)
H5	35853	24948 (69.6)	16176 (45.1)	7228 (20.2)
H6	48358	34961 (72.3)	23150 (47.9)	9842 (20.4)
Total	536224	371660 (69.3)	237608 (44.3)	109563 (20.4)

Table 1: Full cohort caracteristics. The number in parentesis refers to the percentage of uncensored observations in each group.

Variable	Total	Crude death (%)	Cause-spec death(%)	Repeated infarction(%)
Age				
40-44	349	86 (24.6)	48 (13.8)	112 (32.1)
45-49	971	222 (22.9)	121 (12.5)	280 (28.8)
50-54	1845	507 (27.5)	304 (16.5)	458 (24.8)
55-59	2204	637 (28.9)	398 (18.1)	489 (22.2)
60-64	1832	552 (30.1)	369 (20.1)	243 (13.3)
65-69	1136	372 (32.8)	278 (24.5)	121 (10.7)
70-74	313	74 (23.6)	59 (18.9)	36 (11.5)
Year				
87-91	1093	353 (32.3)	155 (14.2)	475 (43.5)
92-96	2176	736 (25.2)	436 (20.0)	633 (29.1)
97-01	2399	698 (29.1)	482 (20.1)	382 (15.9)
02-06	2982	663 (22.2)	504 (16.9)	249 (8.35)
Sex				
male	6448	1862 (28.9)	1211 (18.8)	1366 (21.2)
female	2202	588 (26.7)	366 (16.6)	373 (16.9)
Region				
H1	904	276 (30.5)	198 (21.9)	166 (18.3)
H2	1000	272 (27.2)	161 (16.1)	206 (20.6)
H3	3310	930 (28.1)	584 (17.6)	675 (20.4)
H4	1770	477 (27.0)	324 (18.3)	350 (19.8)
H5	939	281 (29.4)	124 (17.1)	148 (20.4)
H6	939	281 (29.9)	186 (19.8)	194 (20.7)
Total	8650	2450 (28.3)	1577 (18.2)	1739 (20.1)

Table 2: Sibling cohort caracteristics. The number in parentesis refers to the percentage of uncensored observations in each group.

3 Method

To adress the question of heritable prognosis, we consider three events after myocardial infarction: overall mortality, cause specific death and repeated infarction defined in the previous section. To model the time to event we are using the Cox proportional hazards model, which we present in this section, together with the variable selection procedure that we use to fit our models.

3.1 Cox proportional hazards model

Assume we have n independent observations of time, each observation is either censored or noncensored. An observation is uncensored if the event of interest occurs during the time of the study and censored otherwise.

The Cox proportional hazards model is defined as:

$$h_i(t, x, \beta) = \lambda_0(t) \exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}\}$$

where $h_i(t, x, \beta)$ is the hazard function for individual $i = 1, ..., n$ as a function of survival time and subject covariates [3].

The baseline hazard function $\lambda_0(t)$ describes how the hazard changes as a function of survival time and is left unspecified in the Cox model. The baseline hazard function can take any form, except that it cannot be negative. The other expression $\exp\{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}\}$ describes how the hazard changes as a function of the subject covariates. This model was first proposed by Cox in 1972 and is often reffered to as the Cox model [3]. The ratio of the hazard functions for two subjects with covariate vectors $x_0 = (x_{01}, \ldots, x_{0k})$ and $x_1 = (x_{11}, \ldots, x_{1k})$ is

$$HR(t, x_1, x_0) = \frac{\lambda_0(t) e^{x_1 \beta}}{\lambda_0(t) e^{x_0 \beta}} = e^{(x_1 - x_0)\beta} \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_k)^T$ is the column vector of regression coefficients. Suppose covariate number k represents a non-reference cell of a categorical variable and that x_0 and x_1 are identical except for this component, with $x_{0k} = 0$ (variable has not have this cell value) and $x_{1k} = 1$ (variable has this cell value). Then the hazard ratio becomes:

$$\mathrm{HR}(t, x_1, x_0) = e^{(x_1 - x_0)\beta} = e^{(1-0)\beta_k} = e^{\beta_k} \tag{2}$$

and we can construct a 95% confidence interval for β_k by:

$$\exp[\hat{\beta}_k \pm 1.96 \hat{\mathrm{se}}(\hat{\beta}_k)] \tag{3}$$

The hazard ratio plays the same role in interpreting and explaining the results of a survival analysis as the odds ratio for the logistic regression model.

Since the baseline hazard function $\lambda_0(t)$ is not specified in the Cox model it is not possible to maximize the log-likelihood function with respect to β to obtain the maximum likelihood estimator $\hat{\beta}$. Instead we are using the "partial likelihood function" that depends only on the parameter of interest. In absence of censoring, partial likelihood for the multivariate model is

$$L_p(\beta) = \prod_{i=1}^{n} \frac{e^{x_i\beta}}{\sum_{j\in R(t_i)} e^{x_j\beta}}$$

where x_i and t_i is the vector of covariates and time of death for individual i and $R(t_i)$ the set of individuals at risk at time t_i. The log partial likelihood function is:

$$l_p(\beta) = \log L_p(\beta) = \sum_{i=1}^{n} \left\{ x_i\beta - \log \left[\sum_{j\in R(t_i)} e^{x_j\beta} \right] \right\}$$

There are p equations, one for each covariate. To yield the maximum partial likelihood estimators, we set the p equations equal to zero and solve. The equation for the kth covariate is

$$\frac{\delta l_p(\beta)}{\delta \beta_k} = \sum_{i=1}^{n} \left\{ x_{ik} - \frac{\sum_{j\in R(t_i)} x_{jk} e^{x_j\beta}}{\sum_{j\in R(t_i)} e^{x_j\beta}} \right\}$$

The estimator of the covariance matrix of the maximum partial likelihood estimator is the inverse of the observed information matrix evaluated at the maximum partial likelihood estimator,

$$\widehat{\mathrm{Var}}(\hat{\beta}) = I(\hat{\beta})^{-1}$$

where

$$I(\hat{\beta}) = -\frac{\delta^2 l_p(\hat{\beta})}{\delta \hat{\beta}^2}$$

is the observed Fisher information matrix ([3]).

3.2 The proportional hazards assumption

One important property of the Cox model is that the ratio of the hazards is constant over time. Taking the ratio of the hazards for two individuals i and j in equation (1) we see that $\lambda_0(t)$ cancels out.

Assume that we have p covariates and that the n independent observations of time, covariates and censoring indicator are denoted by the triplet (t_i, x_i, c_i), $i = 1, 2, ..., n$ where $c_i = 1$ for uncensored observations and zero otherwise. The Schoenfeld residuals are defined for each covariate but only at uncensored survival times (for censored observations they are set as missing) and are based on the first derivative of the log-likelihood function. The estimator of the Schoenfeld residual for the ith subject on the kth covariate is:

$$\hat{r}_{ik} = c_i(x_{ik} - \hat{\bar{x}}_{w_ik})$$

where

$$\hat{\bar{x}}_{w_ik} = \frac{\sum_{j\in R(t_i)} x_{jk}e^{x_j\hat{\beta}}}{\sum_{j\in R(t_i)} e^{x_j\hat{\beta}}}$$

and $\hat{\beta}$ is the maximum partial likelihood estimator of β. The summation is over all subjects at risk at time t_i denoted $R(t_i)$. The scaled Schoenfeld residuals have better diagnostic power and are therefore used more often in assessing the proportional hazards assumption. The vector of scaled Schoenfeld residuals is the product of the inverse of the covariance matrix times the vector of residuals:

$$\hat{r}^*_{ij} = [\widehat{\text{Var}}(\hat{r}_{ij})]^{-1}\hat{r_{ij}} = r^*_{i1},\ldots,r^*_{ik}$$

The violation of the proportional hazards assumptions is equivalent to interactions between one or more covariates and time. We consider an alternative to the Cox model that has the following specific form of time-varying regression coefficient:

$$\beta_j(t) = \beta_j + \gamma_j \log(t) \tag{4}$$

Under this model, the scaled Schoenfeld residuals have, for the jth covariate, a mean at time t of approximately

$$E[\hat{r}^*_{ij}] \approx \gamma_j \log(t)$$

see [3] for details. To assess the proportional hazards assumption we plot the scaled and smoothed Schoenfeld residuals obtained from the model. If the proportional hazards assumption is fullfilled the average size of the Schoenfeld residuals is independent of time. A smoothed plot of the magnitude of the residual components that shows a relationship with time is therefore an indication of that this assumption is violated.

3.3 Variable selection procedure

For the variable selection procedure we use the "purposeful selection of covariates"-strategy suggested by Hosmer and Lemeshow (1999) [3]. The variable selection procedure begins with a univariate analysis of the association between survival time and all variables that we are considering for the model. To get an overview of how different variables effect the survival time, we start by looking at the Kaplan-Meier estimates of the survival function stratifying on each variable. We plot the Kaplan-Meier curves and use the log rank statistic to test the null hypothesis that there is no difference between survival groups.

We then fit a multivariate model with the variables that are significant at 20-25% level in the univariate analysis. We use the p-values from the Wald test of the individual coefficients to identify variables that could be omitted from the model. We omit the non significant variables one at the time and look at the p-value of the partial likelihood ratio test to confirm that the variable is not significant. Fitting the reduced model it is also important to check if the estimated coefficients change dramatically. This procedure continues until we cannot omit any more variables.

The nex step is to determine if interaction terms are needed in the model. We will confine ourselves to studying the interaction between age and sex, which is interesting from an epidemiological point of view.

3.4 Adressing the question of heritable prognosis

As mentioned above we are mainly interested in assessing the heritability of survival time after acute myocardial infarction. To address this question we identify pairs of siblings where both have suffered from an acute myocardial infarction. We will refer to the siblings in each pair as “first sibling” or “second sibling” depending on if the sibling is first or second (in that pair and in calendar time) to suffer an acute myocardial infarction. We will denote the first sibling in a pair s_{i1} and the second sibling in the same pair s_{i2}. Where i is the siblings pair number, $i = 1, \ldots, n_s$.

Three different outcomes are considered: overall mortality, cause-specific death or repeated infarction. In the first step of the analysis a Cox proportional hazards model is fitted to the whole cohort after excluding the “second siblings” s_{i2}, $i = 1, \ldots, n_s$. We will denote this baseline model m_{1j} for the jth outcome, where $j = 1, 2, 3$ are the three outcomes defined in section 2 (1= overall mortality, 2= cause-specific death and 3=repeated infarction).

We follow the variable selection procedure described in the previous section. We want to capture as much variation in the data as possible that can be explained with the variables that we can adjust for. Therefore we are not very strict about the interpretability of the models m_{1j} and we will also allow for more generous significance levels in the variable selection process if necessary.

From model m_{1j} we calculate the deviance residuals for s_{i1}, $i = 1 \ldots, n_s$ and create a new variable called 'sib prognosis'. The deviance residuals for the ith individual is

$$d_i = \mathrm{sign}(\hat{M_i}) * \sqrt{2[-\hat{M_i} - c_i \log(c_i - \hat{M_i})]}$$

where M_i is the martingale residual for the ith individual

$$M_i = c_i - H(t_i, x_i, \beta)$$

and $c_i = 1$ if the observed survival time t_i is uncensored and zero otherwise and $H(t_i, x_i, \beta)$ is the cumulative hazard at time t_i [9]. The deviance residuals are symmetrically distributed around zero when the fitted model is adequate.

Deviance residuals are positive for individuals who survive for a shorter period than expected and negative for those who survive for a longer time.

We use the deviance residuals quartiles to categorize the siblings into three groups based on their relative survival: “better”, “worse” or “expected”. A sibling pertains to the category “better” if its residual value falls beneath the lowest quartile limit and pertains to the category “worse” if its residual value falls beyond the highest quartile limit. If the residual value falls between these two limits, the sibling pertains to the category “expected”.

To assess the association between the siblings' survival times (within pairs), i.e. a measure of the heritability we fit a Cox proportional hazards model for s_{i2}, $i = 1, \ldots, n_s$, adjusting for necessary variables and with the residuals (categorized) from s_{i1}, $i = 1, \ldots, n_s$ as exposure. This model is denoted m_{2j} for the jth outcome, $j = 1, 2, 3$. Model m_{2j} is developed independently of model m_{1j}, i.e. we will follow the variable selection procedure without regard taken to which variables are included in m_{1j}, with the only difference that we are particularly interested in the association between the 'sib prognosis' and survival time.

3.5 Software

Statistical analysis was performed with the SAS 9.2 software. The procedure *phreg* is used to fit the Cox proportional hazards models. This procedure uses the partial likelihood method to estimate the parameters. The scaled and smoothed Schoenfeld residuals plots are created in R using *cox.zph* in the survival package.

4 Models and Results

4.1 Results from the assessment of the proportional hazards assumption

The scaled and smoothed Schoenfeld residuals are plotted for each variable that we are considering in the model. The plots gives an estimate of the time-dependent coefficient $\beta_j(t)$ defined in (4) (dotted blue line). The proportional hazard assumption is true if $\beta_j(t)$ is a horizontal line. For the outcome overall mortality and cause-specific death the Schoenfeld residuals plots show that the proportional hazards assumption is obviously violated, see Figure 4 and 6 in appendix. For the outcome repeated infarction the assumption about proportional hazards is approximately fulfilled for all variables except for the variable year, see Figure 8 in appendix.

There are different approaches for handling non-proportional hazards. According to our data, over 26% of the individuals that suffer from acute myocardial infarction die the same day. Therefor we partition the time axis into $t = 0$ and $t > 0$ for the outcomes overall mortality and cause specific death i.e. into immediate deaths and non-immediate deaths.

This approach requires that we analyze the outcomes immediate death and non-immediate death separately. In this report we are focusing on the analysis for the non-immediate death. Plotting the Schoenfeld residuals for $t > 0$ (non-immediate death), we can conclude that the previously noticed non-proportionality is not that severe anymore, except for the variable year, see Figures 5 and 7 in appendix.

One way to handle the problem with non-proportional effects is to use a model where we stratify on the non-proportional covariates. We fit separate models for each level of the variable year, using the levels "87-91", "92-96", "97-01" and "02-06", under the constraint that the coefficients are equal but the baseline hazard functions λ_0 are not equal between the levels. Plotting the scaled and smoothed Schoenfeld residuals for the siblings' cohorts we conclude that the proportional hazards assumption is fulfilled for the siblings, see Figures 9, 10 and 11.

4.2 Overall mortality

The full cohort contains 390 754 individuals after the exclusion of the immediate deaths. An individual is censored if alive at the end of the study or if emigrated. In total 41.36% of the individuals are censored and 58.64% are reported as dead during the time of the study.

We chose the model m_{11} based on the variable selection procedure and

Schoenfeld residuals plots. The resulting model contains the three main effects age, region and sex, and also the interaction effect age*sex. Because of violation of the proportional hazards assumption in variable year, we use the stratified model for this variable, as described above. The estimates of the regression coefficients for this model and their p-values are shown in Table 7.

The next step is to fit the model m_{21} for the second siblings with the residuals from the first sibling as exposure. The quartiles are shown in Table 3. The sibling cohort contains 3342 individuals and 86.3% are censored.

Outcome	Lower Quartile	Median	Upper Quartile
Overall mortality	-0.798	-0.622	-0.357
Cause-specific death	-0.492	-0.402	-0.317
Repeated infarction	-0.824	-0.594	0.850

Table 3: Quartiles based on siblings deviance from baseline model

We begin the analysis by considering the univariate analysis of the association between sibprognosis and survival time. In this model sibprognosis worse is highly significant and sibprognosis better is on the borderline of the significance level 5%, as shown in Table 4. Both levels have a survival disadvantages compared to sibprognosis expected.

The adjusted model presented in Table 4 is chosen after following the variable selection procedure presented in section 3.3. After adjusting for age, sibprognosis better is no longer significant at the 5% significance level. The group of siblings with a relatively shorter survival time based on his/hers combination of covariates has a survival disadvantage compared to those with a full sibling with an expected survival time. Age has the strongest effect on the survival time.

4.3 Cause-specific death

The full cohort contains 390 754 observations after exclusion of the immediate death and 75.3 % are censored i.e. alive, dead from other causes than myocardial infarction, alive at the end of the study or emigrated during the time of the study. The estimates of the regression coefficients and their p-values for model m_{12} are showed in Table 8. The sibling cohort contains 3342 observations and 94.4% are censored. The quartiles that we use for categorizing the deviance residuals for the variable sibprognosis are shown in Table 3.

We begin the analysis by considering the univariate analysis of the association between sibprognosis and survival time. The results are presented

Variable	Level	HR	95% CI	p-value
Unadjusted model				
sib prognosis	worse	1.47	(1.17,1.85)	0.0009
	better	1.25	(1.00,1.56)	0.0502
	expected	1	.	.
Adjusted model				
sib prognosis	worse	1.47	(1.17,1.84)	0.0010
	better	1.24	(0.99,1.54)	0.0626
	expected	1	.	.
age	40-49	0.29	(0.19,0.45)	< 0.0001
	50-59	0.59	(0.48,0.72)	< 0.0001
	60-	1	.	.

Table 4: Outcome overall mortality: Estimated HR calculated from (2) and 95% confidence interval calculated from (3) for the unadjusted and adjusted model m_{21}

in Table 5. There is no evidence of association between the sibling's relative survival and the survival time of the second sibling. The adjusted model presented in Table 5 is chosen after following the variable selection procedure presented in section 3.3. Age has the strongest effect on survival time.

Variable	Level	HR	95% CI	p-value
Unadjusted model				
sib prognosis	worse	0.99	(0.52,1.17)	0.2268
	better	1.00	(0.72,1.39)	0.9911
	expected	1	.	.
Adjusted model				
sib prognosis	worse	0.83	(0.55,1.24)	0.3639
	better	1.01	(0.73,1.40)	0.9428
	expected	1	.	.
age	40-49	0.12	(0.04,0.39)	0.0004
	50-59	0.70	(0.52,0.95)	0.0214
	60-	1	.	.

Table 5: Outcome cause-specific death: Estimated HR calculated from (2) and 95% confidence interval calculated from (3) for the unadjusted and adjusted model m_{22}

4.4 Repeated infarction

For the outcome repeated infarction the cohort contains 330 612 individuals and 67% of those are censored. To be included in this cohort the individual needs to be at risk for a second myocardial infarction i.e. the individual must be alive 28 days after the index event. The estimates of the regression coefficients for the model m_{13} and their p-values are shown in Table 9.

The sibling cohort contains 3037 individuals and 85% of those were censored. The quartiles that we use for categorizing the deviance residuals for the variable sibprognosis are shown in Table 3. We begin the analysis by considering the univariate analysis of the association between sibprognosis and survival time. The results are presented in Table 6. There is no evidence of association between the siblings relative survival (time to repeated infarction) and the survival time for the second sibling.

The adjusted model presented in Table 6 is chosen after following the variable selection procedure presented in section 3.3. Note that proportional hazards assumption is fullfilled for the variable year in the model based on s_{i2}, $i = 1, ..., 3037$ and there is no need for stratifying on this variable here. Adjusting for age and year there is still no evidence of association between the siblings relative survival (time to repeated infarction) and the survival time for the second sibling.

Variable	Level	HR	95% CI	p-value
Unadjusted mode				l
sib prognosis	worse	1.11	(0.89,1.39)	0.3677
	better	0.98	(0.78,1.23)	0.8615
	expected	1	.	.
Adjusted model				
sib prognosis	worse	1.03	(0.82,1.30)	0.7965
	better	0.86	(0.68,1.09)	0.2157
	expected	1	.	.
age	40-49	0.61	(0.43,0.88)	0.0072
	50-59	0.81	(0.65,1.00)	0.0460
	60-	1	.	.
year	87-91	1.92	(1.23,3.00)	0.0040
	92-96	1.65	(1.29,2.13)	< 0.0001
	02-06	0.91	(0.71,1.17)	0.4711
	97-01	1	.	.

Table 6: Outcome repeated infarction: Estimated HR calculated from (2) and 95% confidence interval calculated from (3) for the unadjusted and adjusted model m_{23}

5 Discussion

The aim of this study was to assess the heritability of survival time after acute myocardial infarction. We have studied three different outcomes of myocardial infarction, overall mortality, cause-specific death and repeated infarction. For the outcome overall mortality the results from our analysis indicates that their is an association between full siblings survival time after adjusting for age, sex, year and county. For the outcome cause specific death and repeated infarction we cannot see any evidence for association between full siblings survival experience.

When a death is reported to the Cause of death register there can be as many as 20 different causes reported as contributing to the death in excess of the underlying cause of death. We might loose some important information about the cause of death since we only have considered the underlying cause of death to identify the cause-specific deaths. Because of the design of the analysis, the model is sensitive to this kind of measurement errors. The misclassification that might occur afflicts both the exposure (the outcome for s_{i1}) and the outcome (the outcome for s_{i2}).

We see an association between full sibling survival for the outcome overall mortality but not for the outcome cause-specific death. This indicates that there might be another explanation to the co-morbidity between full siblings that is not due to the myocardial infarction incidence. The full siblings might have shared frailties i.e. diseases etc. that confound our results. As a continuation to this study it would be interesting to add information about various factors that may influence health and subsequent risk of death such as factors related to lifestyle (e.g. smoking, diet, lack of exercise) and co-morbidities (e.g. diabetes, hypertension, hypercholesterolemia). If we had seen significant heritability the next step would be to study the corresponding results using half siblings, to determine if what we see is a genetic effect or is a consequence of the childhood environment.

Looking at the deviance residuals for the first siblings s_{i1}, $i = 1, ..., n_s$ to suffer from an acute myocardial infarction we see that these are strongly skewed towards negative values for both outcome overall mortality and death caused by myocardial infarction. The siblings are doing better as a group compared to the whole cohort. Possible explanations to this phenomenon can be that the siblings are more likely to have a family history of myocardial infarction, which can result in closer medical supervision, preventive health care and awareness of the disease. We also know that the siblings' cohort is younger than the full cohort. This might be an explanation if the baseline model is not flexible enough to capture the variation due to age.

We have chosen to treat the outcome immediate death and non-immediate deaths as two different outcomes. Medically this makes sense since we know that the first hours after an acute myocardial infarction are critical and that factors such as how fast you get under medical care is crucial for the survival. Statistically, treating immediate death and non-immediate death in the same model has the consequence that a large proportion of the observations are concentrated in the same point $t = 0$. Separating the two outcomes leads to a much more transparent statistical model. The proportional hazards assumption holds approximately after removing the immediate deaths except for the variable year. It is not obvious how to categorize a continuous variable for the stratification purpose, which might lead to bias in the estimated regression coefficients as well as in the residual distribution. Another approach could be model year by means of a smooth spline function.

The focus in this report has been on the Cox proportional hazards model for the first day survivors. However, we have also done some preliminary analysis for the immediate death in a similar manner. We used a logistic regression model with the first sibling outcome as exposure. Preliminary results from this analysis indicates that there is a significant siblings effect, but the model needs further development and examination before one can draw any definite conclusions. This will be subject to future research.

References

[1] Dudas K. et al. (2010). Long-term prognosis after hospital admission for acute myocardial infarction from1987 to 2006, *International Journal of Cardiology*, doi:10.1016/j.ijcard.2010.10.047.

[2] Harpaz D., Behar S. et al. (2004). Family History of Coronary Artery Disease and Prognosis after First Acute Myocardial Infarction in a National Survey, *Cardiology* **102** (2004), 140–146.

[3] Hosmer D.W. JR, Lemeshow S. (1999). *Applied Survival Analysis*, Wiley Series in Probability and Statistics.

[4] Leander K., Wiman B. et al. (2007). Primary risk factors influence of recurrent myocardial infarction/death from coronary heart disease: results from the Stockholm Heart Epidemiology Program (SHEEP), *European Journal of Cardiovascular Prevention and Rehabilitation* **14**, 532–537.

[5] Little, R. J Rubin, D.B. (2002). *Statistical Analysis with Missing Data* Wiley.

[6] MacIntyre K., Stewart S. et al.(2001). Gender and Survival: A Popunlation-Based Study of 201,114 Men and Women Following a First Acute Myocardial Infarction, *Journal of the American College of Cardiology* Vol. 38, No. 3. 729–35.

[7] Rosengren A., Spetz C.-L. et al. (2001). Sex differences in survival after myocardial infarction in Sweden, Data from the Swedish National Acute Myocardial Infarction register, *European Heart Journal* **22**, 314–322.

[8] Selzer A. (1992). *Understanding Heart Disease*, University of California Press.

[9] Therneau T. M. , Grambsch P. M. (2000).*Modeling Survival - Extending the Cox Model*, Springer.

[10] SCB*H-regioner (pdf)* Can be downloaded from `http://www.scb.se/Pages/List____257369.aspx`

[11] *Öppna jämförelser av hälso- och sjukvårdens kvalitet och effektivitet, jämförelser mellan landsting*, Sveriges kommuner och landsting och Socialstyrelsen 2009.

[12] *Folkhälsorapporten 2009*, Socialstyrelsen 2009.

[13] *Hjärtinfarkter 1987-2008 samt utskrivna efter vård för akut hjärtinfarkt 1987-2009*, Socialstyrelsen 2010.

[14] Socialstyrelsen `http://www.socialstyrelsen.se/register`

A Tables and Figures

Parameter	Level	Estimate	Standard Error	p-value
Sex	male	0.118	0.018	< 0.0001
Age	40-44	-3.256	0.074	< 0.0001
	45-49	-3.246	0.051	< 0.0001
	50-54	-3.130	0.039	< 0.0001
	55-59	-2.778	0.028	< 0.0001
	60-64	-2.396	0.020	< 0.0001
	65-69	-1.983	0.016	< 0.0001
	70-74	-1.577	0.014	< 0.0001
	75-79	-1.104	0.013	< 0.0001
	80-84	-0.670	0.012	< 0.0001
	85-89	-0.303	0.013	< 0.0001
Region	H1	-0.090	0.009	< 0.0001
	H2	-0.111	0.009	< 0.0001
	H3	-0.090	0.008	< 0.0001
	H4	-0.066	0.008	< 0.0001
	H5	-0.031	0.011	0.0036
Age*Sex	(40-44)*male	0.316	0.084	0.0002
	(45-49)*male	-0.224	0.058	0.0001
	(50-54)*male	-0.058	0.045	0.1974
	(55-59)*male	-0.081	0.034	0.0166
	(60-64)*male	-0.070	0.027	0.0087
	(65-69)*male	-0.046	0.023	0.0447
	(70-74)*male	-0.007	0.021	0.7398
	(75-79)*male	-0.004	0.020	0.8566
	(80-84)*male	-0.014	0.020	0.4951
	(85-89)*male	0.0002	0.0212	0.9925

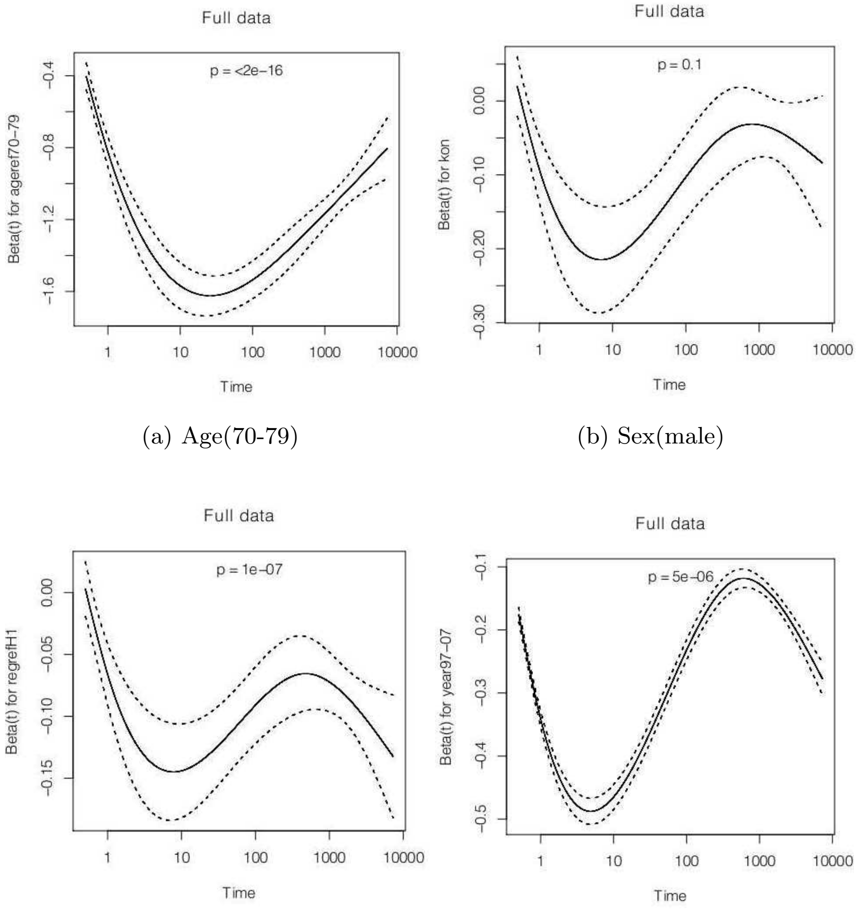
Table 7: Overall mortality: Estimates for the baseline model m_{11}

Parameter	Level	Estimate	Standard Error	p-value
Sex	male	0.154	0.026	< 0.0001
Age	40-44	-3.186	0.140	< 0.0001
	45-49	-2.869	0.082	< 0.0001
	50-54	-2.947	0.068	< 0.0001
	55-59	-2.486	0.045	< 0.0001
	60-64	-2.091	0.032	< 0.0001
	65-69	-1.702	0.025	< 0.0001
	70-74	-1.309	0.021	< 0.0001
	75-79	-0.877	0.019	< 0.0001
	80-84	-0.538	0.018	< 0.0001
	85-89	-0.246	0.019	< 0.0001
Region	H1	-0.144	0.014	< 0.0001
	H2	-0.196	0.014	< 0.0001
	H3	-0.129	0.012	< 0.0001
	H4	-0.086	0.012	< 0.0001
	H5	-0.060	0.016	0.0002
Age *Sex	(40-44)*male	0.074	0.152	0.6246
	(45-49)*male	-0.206	0.093	0.027
	(50-54)*male	0.139	0.076	0.0684
	(55-59)*male	-0.061	0.054	0.2587
	(60-64)*male	-0.085	0.042	0.0408
	(65-69)*male	-0.074	0.034	0.0306
	(70-74)*male	-0.061	0.031	0.0479
	(75-79)*male	-0.085	0.029	0.0041
	(80-84)*male	-0.048	0.029	0.1009
	(85-89)*male	-0.032	0.031	0.3053

Table 8: Cause-specific death: Estimates for the baseline model m_{12}

Parameter	Level	Estimate	Standard Error	p-value
Sex	male	0.154	0.037	< 0.0001
Age	40-44	-1.639	0.074	< 0.0001
	45-49	-1.558	0.052	< 0.0001
	50-54	-1.585	0.043	< 0.0001
	55-59	-1.334	0.035	< 0.0001
	60-64	-1.162	0.030	< 0.0001
	65-69	-0.933	0.026	< 0.0001
	70-74	-0.735	0.025	< 0.0001
	75-79	-0.471	0.024	< 0.0001
	80-84	-0.255	0.024	< 0.0001
	85-89	-0.086	0.025	0.0007
Region	H1	-0.046	0.013	0.0004
	H2	-0.095	0.013	< 0.0001
	H3	-0.026	0.011	0.0202
	H4	-0.015	0.012	0.2168
	H5	-0.020	0.016	0.2072
Age*Sex	(40-44)*male	0.121	0.086	0.1561
	(45-49)*male	0.092	0.063	0.1454
	(50-54)*male	0.145	0.055	0.0084
	(55-59)*male	-0.021	0.045	0.6553
	(60-64)*male	-0.018	0.044	0.6877
	(65-69)*male	-0.070	0.041	0.0905
	(70-74)*male	-0.030	0.040	0.4475
	(75-79)*male	-0.066	0.040	0.0964
	(80-84)*male	-0.032	0.040	0.4199
	(85-89)*male	-0.016	0.042	0.7079

Table 9: Repeated infarction: Estimates for the baseline model m_{13}



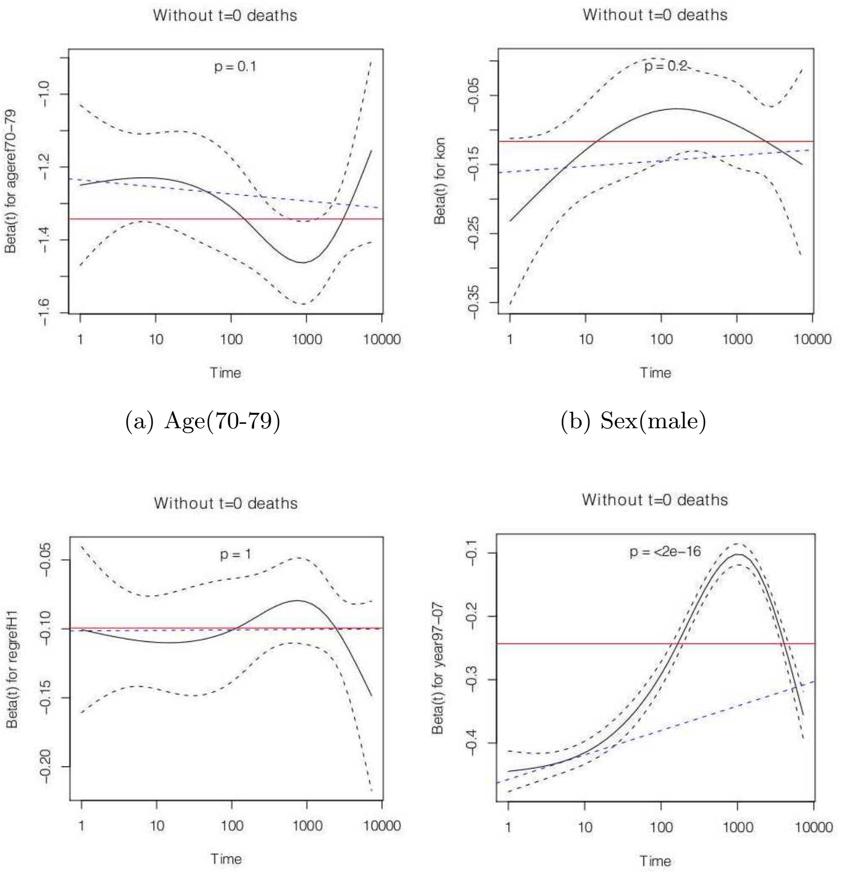
(a) Age(70-79)

(b) Sex(male)

(c) Region(H1))

(d) Year(97-07)

Figure 4: Overall mortality: scaled and smoothed Schoenfeld residuals



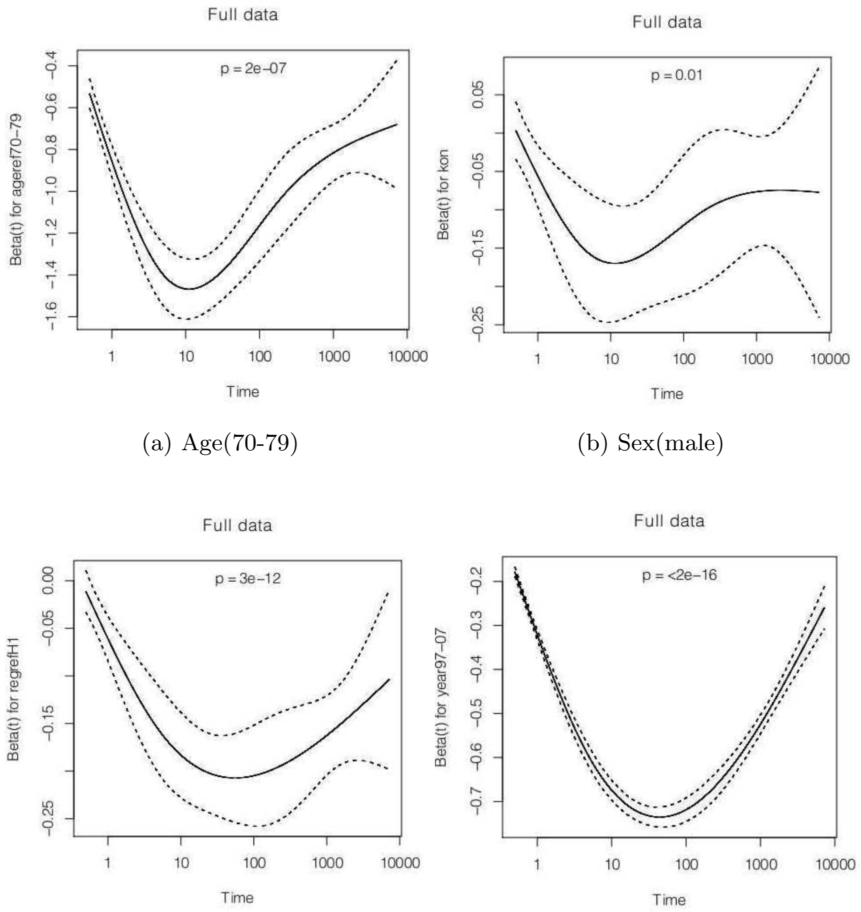
(a) Age(70-79)

(b) Sex(male)

(c) Region(H1))

(d) Year(97-07)

Figure 5: Overall mortality: scaled and smoothed Schoenfeld residuals for $t > 0$



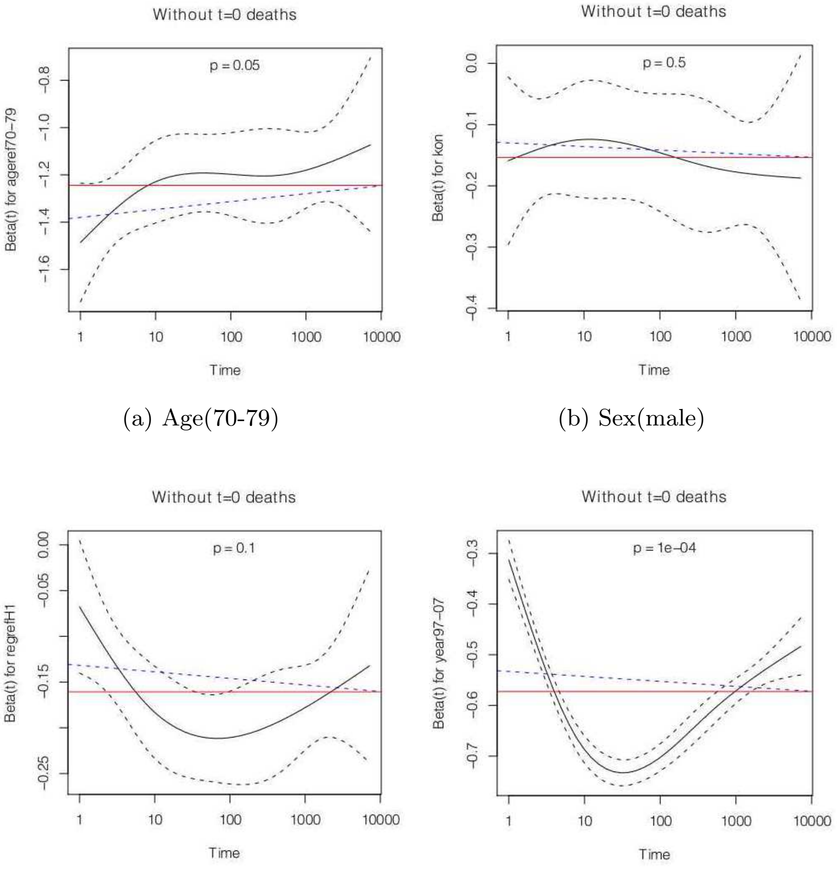
(a) Age(70-79)

(b) Sex(male)

(c) Region(H1))

(d) Year(97-07)

Figure 6: Cause-specific death: scaled and smoothed Schoenfeld residuals



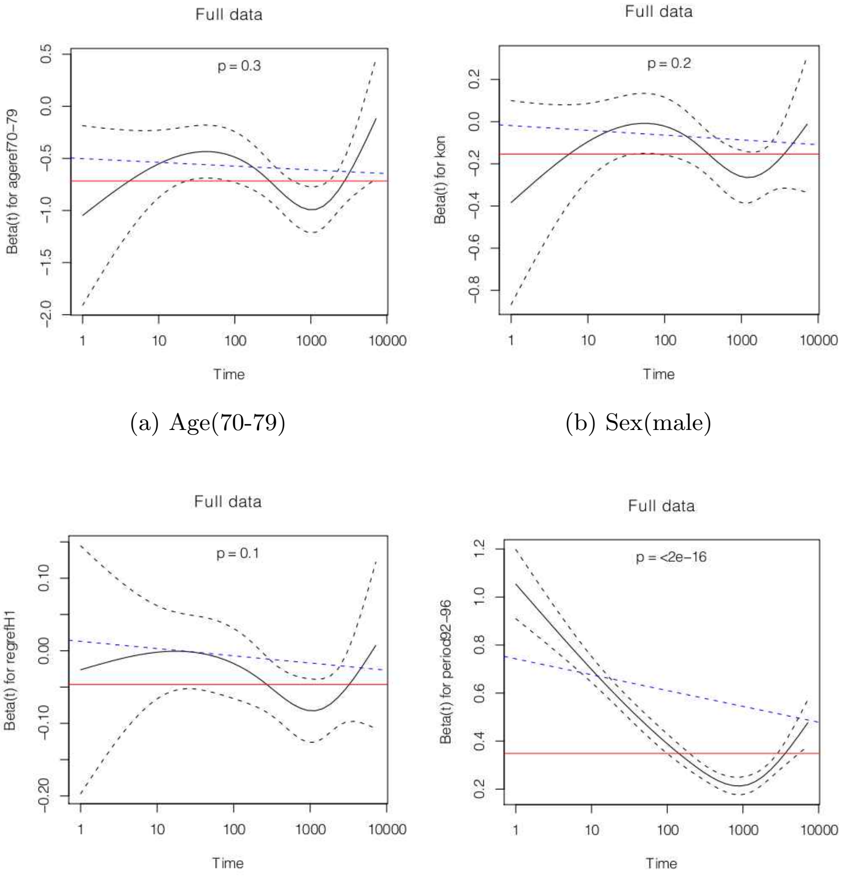
(a) Age(70-79)

(b) Sex(male)

(c) Region(H1))

(d) Year(97-07)

Figure 7: Cause-specific death: scaled and smoothed Schoenfeld residuals for $t > 0$



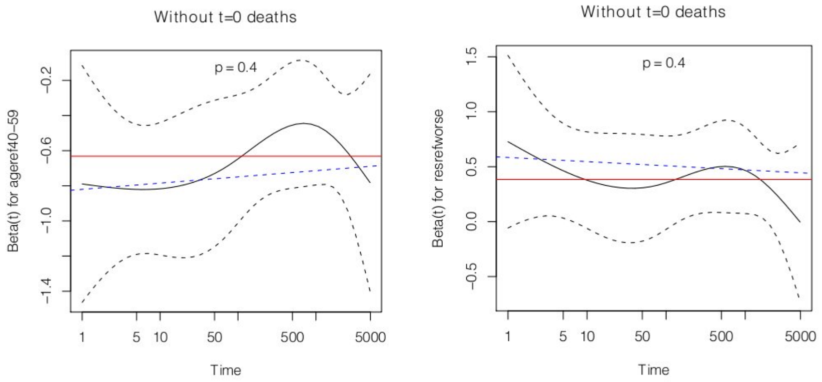
(a) Age(70-79)

(b) Sex(male)

(c) Region(H1))

(d) Year(97-07)

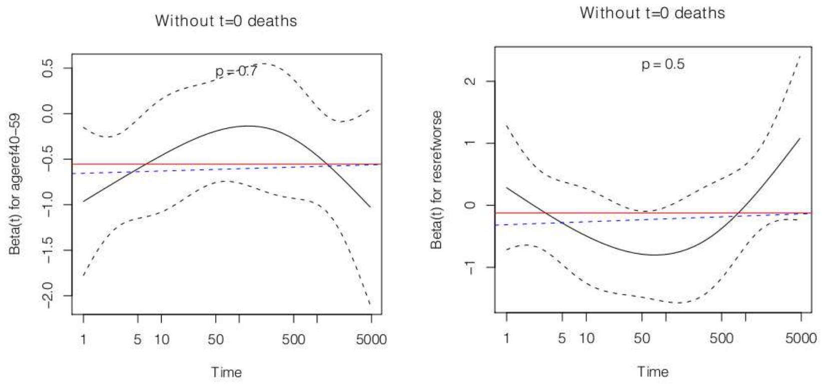
Figure 8: Repeated infarction: scaled and smoothed Schoenfeld residuals



(a) Age(40-59)

(b) Sib prognosis(worse)

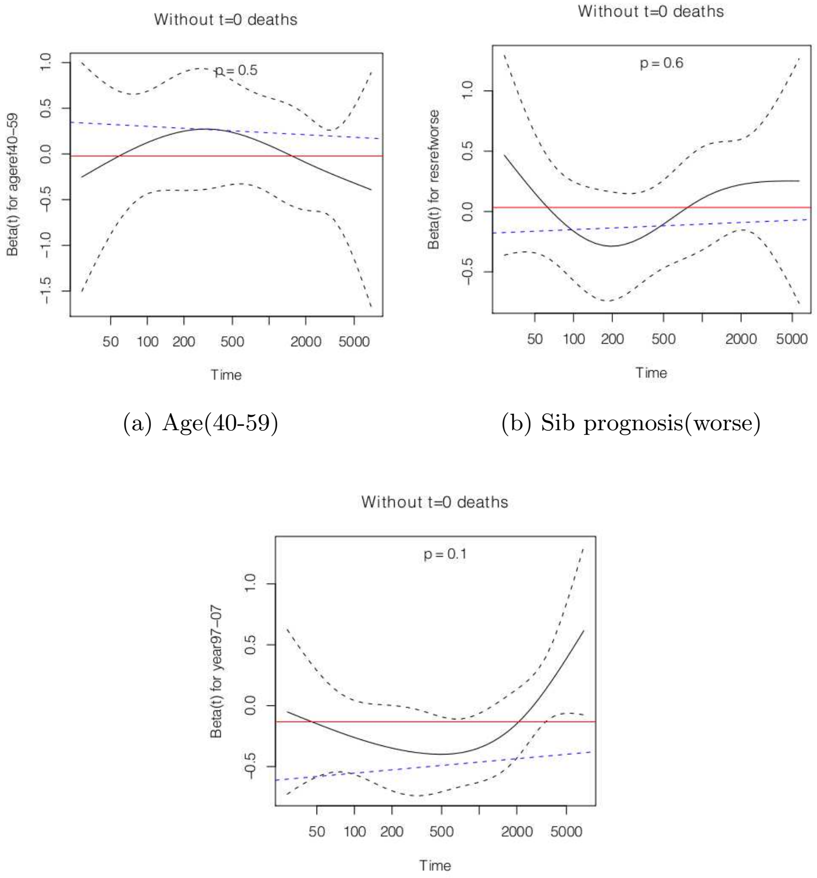
Figure 9: Overall mortality: smoothed Schoenfeld residuals for the siblings



(a) Age(40-59)

(b) Sib prognosis(worse)

Figure 10: Cause-specific death: smoothed Schoenfeld residuals for the siblings



(a) Age(40-59)

(b) Sib prognosis(worse)

(c) Year(97-07)

Figure 11: Repeated infarction: smoothed Schoenfeld residuals for the siblings