NTNU

Norwegian University of
Science and Technology

Risk Factors for Breast, Uterine and Ovarian Cancer: A competing Risks Analysis

Lillian Grude

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Bo Henry Lindqvist, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

• Give an introduction to the theory of competing risks, with emphasis on regression, modeling and inference.

• Use the theory and appropriate computer programs to analyse data from a screening program organized by the Norwegian Cancer Society, studying common risk factors for breast cancer, uterine cancer and ovarian cancer.

Assignment given: 24 January 2011

Supervisor: Bo Henry Linqvist

Preface

This thesis is carried out at the Department of Mathematical Sciences and Technology at the Norwegian University of Science and Technology (NTNU) during the period January 2011 to June 2011.

I owe particular gratitude to Professor Bo Lindqvist who has been my supervisor throughout this thesis. His suggestions, support and encouraging guidance have been of invaluable importance.

I also wish to express my gratitude to Pål Romundstad for an interesting discussion that led to a exciting problem description, and also for access to the dataset.

Furthermore, I want to thank Signe Opdahl for epidemiological advice during this work. A special thanks to her also for her excellent data description which made my work a lot easier.

Sincere thanks is also given to Hilde Grude Borgos for proofreading this report and giving me useful suggestions for language improvements.

Last, but at least, I want to thank all my friends and fellow students at room 393C in "Matteland" for interesting discussions and a friendly working environment, and especially to Karianne Lien for informal small talk and scientific inputs during everyday life.

Trondheim, June 2011

Lillian Grude

Abstract

A competing risks situation arises when a unit can fail due to several distinct failure types. In a competing risk situation, standard techniques from survival analysis may lead to incorrect and biased results. In this thesis, the theory of competing risks is used to identify possible risk factors for breast, uterine and ovarian cancer. This has been done by regression on the cause specific hazard functions, the subdistribution hazard functions and two approximate methods. Cox regression is used for a complete analysis of the medical data.

By following 61457 women over approximately 50 years, it has been observed 3407 cases of breast cancer, 934 of uterine cancer and 843 of ovarian cancer. Summarized, it has been found that several births decrease the risk of breast, uterine and ovarian cancer. Obesity is associated with increasing risk of ovarian cancer for postmenopausal women, but not premenopausal. A long reproductive period (early menarche and/or late menopause) and high BMI increases the risk of breast and uterine cancer. Late first and last birth decreases the risk of uterine cancer, while it increases the risk of breast cancer. The data used in the analysis is selected from a screening program organized by the Norwegian Cancer Society for early diagnosis of breast cancer.

Contents

1 Introduction **1**

2 Theory **3**

2.1 Model specification and mathematical definition 3

2.1.1 Bivariate random variable representation 3

2.1.2 Latent failure time representation 6

2.2 Likelihood function formulation 8

2.3 Nonparametric estimation 9

2.4 Regression models 10

2.4.1 Cause specific hazard functions 10

2.4.2 Cumulative incidence functions 11

2.5 Effect of a covariate 13

2.6 Approximate Cox regression 16

2.6.1 Normally distributed covariates 17

2.6.2 Gamma distributed covariates 19

2.6.3 Binomially distributed covariates 19

2.6.4 Multinomially distributed covariates 20

2.6.5 General distribution for the covariates 21

2.6.6 Local estimation of β_d 22

2.7 Approximate Fine and Grey regression 22

3 Data description **25**

3.1 Material 25

3.2 Study variables 26

3.2.1 Age 26

3.2.2 Reproductive variables 27

3.2.3 Demographic variables 28

3.2.4 Height and weight 29

3.2.5 Reproductive period 30

3.2.6 Lactation 31

3.2.7 Abortion 31

4 Competing risks **33**
4.1 Breast cancer . . . 33
4.1.1 General information about breast cancer . . . 33
4.1.2 Current knowledge . . . 35
4.1.3 Breast cancer in HUNT0 . . . 36
4.2 Cancer of the Uterus . . . 37
4.2.1 General information about uterine cancer . . . 37
4.2.2 Current knowledge . . . 39
4.2.3 Uterine cancer in HUNT0 . . . 39
4.3 Ovarian cancer . . . 40
4.3.1 General information about ovarian cancer . . . 40
4.3.2 Current knowledge . . . 41
4.3.3 Ovarian cancer in HUNT0 . . . 42

5 Explanatory data analysis **43**
5.1 Parity . . . 43
5.2 BMI . . . 45
5.3 Age at first birth . . . 49
5.4 Age at last birth . . . 49
5.5 Age at menarche . . . 51
5.6 Age at menopause . . . 52
5.7 Other . . . 53

6 Regression on the cause specific hazard functions **55**
6.1 Model description . . . 55

7 Medical results **59**
7.1 Breast cancer . . . 59
7.2 Uterine cancer . . . 62
7.3 Ovarian cancer . . . 65

8 Concluding remarks **69**

References **73**

Chapter 1

Introduction

A competing risks situation arises when the unit under study can experience any one out of several distinct failure types, and the occurrence of one precludes the occurrence of a competing event. Competing risks have many similarities with survival analysis, which involves modeling the time span from a given time origin until the occurrence of one single type of event. However, the theory of competing risks does not originate from survival analysis. The theory can be tracked back to 1760, when David Bernoulli studied possible consequences of eradication of smallpox on mortality rates.

The definition of competing risks can be illustrated by an example from the medical field where it arises naturally. Death due to heart failure may be the cause of interest, death due to other causes will then be competing events because they prevent death due to heart-failure. The time and cause of death is observed, while no information is obtained regarding the diseases that did not cause death. The competing risks framework also includes settings where different possible events are not mutually exclusive but the interest lies on the first occurring event, as for example the risk of getting a certain cancer or failure of a component in a system in industrial reliability testing.

A competing risks model can be expressed graphically with an initial state and p different endpoints, see Figure 1.1.

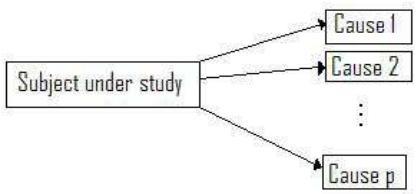


Figure 1.1: Competing risks situation with p causes of failure.

In this thesis, the theory of competing risks is used to analyse data from a screening program organized by the Norwegian Cancer Society for early diagnosis of breast cancer. The selected competing events are breast cancer, uterine cancer and ovarian cancer. These are cancers with assumed similar risk factors, and the data set is therefore appropriate for this analysis. From 01.01.1961 to 15.02.2010, a period of approximately 50 years, 3407 out of 61457 women were diagnosed with breast cancer, 934 with uterine cancer and 843 with ovarian cancer.

To study the relationship between breast, uterine and ovarian cancer risk and selected explanatory variables, regression on the cause specific hazard function and regression on the subdistribution hazard function will be performed. Four methods will be used, Cox regression, regression on the subdistribution hazard functions, Approximate Cox regression and Approximate Fine and Grey regression. The latter two methods are described in detail in the theory. Out of the four methods, the most adequate are used for a comprehensive analysis of cancer risk in relation to selected explanatory variables.

This thesis is divided into eight chapters. Chapter 2 contains general theory about competing risks, where the main focus is on the bivariate random variable representation of competing risks. The chapter also discusses two approximate methods, Approximate Cox regression and Approximate Fine & Grey regression. Chapter 3 gives an introduction to the dataset analyzed in this thesis. A description of the competing events is given in Chapter 4. Explanatory data analysis is found in Chapter 5. Chapter 6 explains the mathematical model that best describes the relationship between the competing events and the explanatory variables. Chapter 7 contains medical results. A conclusion is found in Chapter 8.

Chapter 2

Theory

2.1 Model specification and mathematical definition

There are two different mathematical definitions of competing risks, the first is related to the joint distribution of time and cause of failure and the second relies on p hypothetical failure times. The theory in this section is selected from Lindqvist [24], Prentice et al. [30] and M.J. Crowder [7]. The first part of this chapter (Section 2.1 - 2.5) is an extension of the work done in the specialization project by Grude [12].

2.1.1 Bivariate random variable representation

For each subject the pair (T, D) is observed, where $T \geq 0$ is the time of failure and $D \in \{1, 2, ..., p\}$ is the failure cause. T is assumed to be a continuous and positive random variable while D belongs to exactly one of p different failure types. If an event of type d occurs first, $D = d$, T is then the time at which this event occurred.

The joint distribution between T and D is completely specified by either the cumulative incidence functions, $F_d(t)$, or the cause specific hazard functions, $\lambda_d(t)$.

The cumulative incidence functions, CIF, for a failure of type d is defined by

$$F_d(t) = P(T \leq t, D = d), \qquad t > 0, \qquad d \in \{1, 2, ..., p\},$$

and corresponds to the sub-distribution function for the probability of failure from cause d in the presence of the competing events. The cumulative incidence function is also known as the marginal probability function and the crude incidence.

The marginal distribution function of T is the sum of the cumulative incidence functions for all failure types, i.e. the probability of failure from any type of event

at or before time t,

$$F(t) = P(T \leq t) = \sum_{d=1}^{p} F_d(t).$$

Equivalently, the marginal distribution function of T can be described by the overall survival function, which is the probability of surviving from all failure types up to time t,

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

The sub-survival function for cause d represents the probability of not failing from cause d before time t,

$$\bar{F}_d(t) = P(T > t, D = d), \qquad d \in \{1, 2, ..., p\}.$$

The sub-survival function is not a proper survivor function because

$$P(D = d) = F_d(\infty) = \bar{F}_d(0) = p_d, \qquad d \in \{1, 2, ..., p\},$$

which is strictly below 1 if there are at least two competing events. In the above equation, p_d is the marginal probability for cause D.

The cumulative incidence function and the sub-survival function are related by

$$F_d(t) + \bar{F}_d(t) = p_d, \qquad d \in \{1, 2, ..., p\},$$

hence, the sub survivor function is the complementary of the probability of failing from cause d, $P(D = d)$.

The sub-density functions $f_d(t)$ from cause d, when they exists, are given by

$$f_d(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = d)}{dt} = \frac{d}{dt} F_d(t), \qquad d \in \{1, 2, ..., p\}. \tag{2.1}$$

The cause specific hazard function represents the instantaneous failure rate for cause d and is given by

$$\begin{aligned} \lambda_d(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d)}{\Delta t P(T > t)} \\ &= \frac{f_d(t)}{\bar{F}(t)}, \qquad d \in \{1, 2, ..., p\}. \end{aligned} \tag{2.2}$$

The cause specific hazard functions can be seen as the fundamental concept in competing risks and are also referred to as the sub-hazard functions or mode-specific

hazard functions. Because the cumulative incidence function is a joint distribution, the relationship between the various subfunctions is not as expected from standard survival analysis with one single endpoint, i.e. $\lambda_d(t) \neq f_d(t)/\bar{F}_d(t)$ in general.

The hazard function of T is defined as the sum of the cause specific hazard functions and can be interpreted as the instantaneous failure rate from any cause,

$$\lambda(t) = \sum_{d=1}^{p} \lambda_d(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.3}$$

The relationship between the cumulative incidence functions and the cause specific hazard rates follows from equation (2.1) and (2.2),

$$F_d(t) = \int_0^t f_d(u)du = \int_0^t \lambda_d(u)\bar{F}(u)du, \qquad d \in \{1, 2, ..., p\}. \tag{2.4}$$

By integrating the hazard function, the cumulative cause specific hazard function is obtained,

$$\Lambda_d(t) = \int_0^t \lambda_d(u)du, \qquad d \in \{1, 2, ..., p\}.$$

From equation (2.3) it follows that the cumulative hazard function of T is $\Lambda(t) = \int_0^t \lambda(u)du = \sum_{d=1}^{p} \Lambda_d(t)$, so the overall survival function is given as

$$\bar{F}(t) = e^{-\Lambda(t)} = e^{-\sum_{d=1}^{p} \Lambda_d(t)} = \prod_{d=1}^{p} \bar{G}_d^*(t), \qquad \text{where } \bar{G}_d^*(t) = e^{-\Lambda_d(t)}. \tag{2.5}$$

It is important to notice that although $\bar{G}_d^*(t)$ is identifiable from the joint distribution of (T, D), it should not be interpreted as a survival function because it is not in general the distribution of any observable random variable. With independent latent failure times, $\bar{G}_d^*(t)$ can be interpreted as a marginal distribution of the latent failure times, T_d (see section 2.1.2).

In 1988 Grey [11] desired (of reasons to be given in Section 2.5) a model on the form $F_d(t) = 1 - e^{-\int_0^t w_d(u)du} = 1 - e^{-W_d(u)}$, and introduced the subdistribution

hazard functions

$$\begin{aligned}
w_d(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d | T > t \cup (T \leq t \cap D \neq d))}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d \cup [T > t \cup (T \leq t \cap D \neq d)])}{\Delta t P(T > t \cup (T \leq t \cap D \neq d)} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d \cup [T > t \cup (T \leq t \cap D \neq d)])}{\Delta t P(T > t \cup D \neq d)} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = d)}{\Delta t (1 - P(T \leq t, D = d))} \\
&= \frac{dF_d(t)/dt}{1 - F_d(t)} \\
&= \frac{f_d(t)}{\bar{F}_d(t)} \\
&= \frac{-d(\log(1 - F_d(t))}{dt}, \qquad\qquad d \in \{1, 2, ..., p\}. \qquad (2.6)
\end{aligned}$$

The subdistribution hazard function is the hazard corresponding to the cumulative incidence function. The interpretation of the subdistribution hazard function is unnatural. It is the instantaneous failure rate from cause d given that a subject either has survived or has already failed due to a competing event. The subdistribution hazard can be thought of as the hazard function for the improper random variable $T^* = I(D = d) \times T + I(D \neq d) \times \infty$, because $P(T^* \leq t) = P(T \leq t, D = d)$.

The relationship between the cumulative incidence functions and the subdistribution hazard functions follows from equation (2.6),

$$F_d(t) = 1 - e^{-\int_0^t w_d(u)du}, \qquad d \in \{1, 2, ..., p\}. \qquad (2.7)$$

2.1.2 Latent failure time representation

The second definition of competing risks assumes that there is an associated non-negative failure time, $T_1, T_2, ..., T_p$, to each competing risk. Among the p different competing risks the failure time of the first event is observed and the other failure times are latent. Once the system has failed, the remaining lifetimes are lost to observation, i.e. we only observe the pair (T, D) where $T = \min\{T_1, ..., T_p\}$ is the time to the first failure and $D = \{d; T_d \leq T_p \forall p\}$ is the cause of the first failure. It is assumed that ties cannot occur, $\mathrm{P}(T_p \neq T_d) = 0 \ \forall p \neq d$.

The joint survival function for the latent failure times, $T_1, ...T_p$, is described by

$$\bar{K}(t_1, ..., t_p) = P(T_1 > t_1, ..., T_p > t_p).$$

It follows that the sub-density functions are

$$f_d(t) = -\Big(\frac{\partial \bar{K}(t_1, ..., t_p)}{\partial t_d}\Big)_{t_1=t_2=...=t_p=t}, \qquad d \in \{1, 2, ..., p\}.$$

The survival function of T is the probability of survival up to time t. This means that all the potential failure times have to exceed t,

$$\bar{F}(t) = \bar{K}(t, t, ..., t).$$

The cause specific hazard functions are given as

$$\begin{aligned}\lambda_d(t) &= \lim_{\Delta t \to 0} \frac{P(t < T_d \leq t + \Delta t | T > t)}{\Delta t} \\ &= -\Big(\frac{\partial \log \bar{K}(t_1, ..., t_p)}{\partial t_d}\Big)_{t_1=t_2=...=t_p=t}, \qquad d \in \{1, 2, ..., p\},\end{aligned}$$

and the marginal distribution of T_d as

$$\bar{G}_d(t) = P(T_d > t) = \bar{K}(t_1 = 0, t_2 = 0, ..., t_d = t, ..., t_p = 0), \qquad d \in \{1, 2, ..., p\},$$

with corresponding hazard rates,

$$h_d(t) = \frac{-\bar{G}_d'(t)}{\bar{G}_d(t)} = -\Big(\frac{\partial \log \bar{K}(t_1, ..., t_p)}{\partial t_d}\Big)_{t_d=t, t_p=0, p\neq d}, \qquad d \in \{1, 2, ..., p\}. \tag{2.8}$$

Only quantities that can be expressed in terms of the cause specific hazard functions are identifiable. This will be discussed further in Section 2.2. The marginal distribution of the latent failure times cannot be expressed in terms of the cause specific hazard functions without additional assumptions, they are therefore non-identifiable.

The identifiability problem addresses the difficulty of finding the joint and even the marginal distribution of the latent failure times from observations of (T, D). Tsiatis [34], among other, noted that there are several different joint distributions of the latent failure times that gives the same distribution of (T, D). Further, there exists a unique model with independent risks that gives the same cumulative incidence function $F_d(t)$ as the model with dependent risks would give. This model is defined by

$$\bar{K}(t_1, ..., t_p) = \prod_{d=1}^{p} \bar{G}_d^*(t),$$

where $\bar{G}_d^*(t)$ are given by equation (2.5). The model with independent risks and any compatible model with dependent risks fit the data equally well, and it is therefore not possible to know which model that is correct by observing (T, D).

There are several ways to deal with the identifiability problem, one is to assume independent risks. If the failure times T_d are independent, the marginal distribution of T_d is identifiable, and the cause specific hazard functions, $\lambda_d(t)$ given in (2.2), and the hazard of the marginal distribution, $h_d(t)$ given in (2.8), are equal. However, the hypothesis of independence cannot be tested from the data of the form (T, D). Another way to deal with the identifiable problem is to assume a known copula for the latent variables. Zheng and Klein [35] proved that the marginal distribution are estimable from observations (T, D) when the dependence is given in the model, by a known copula. Peterson [28] dealt with the identifiability problem by introducing bounds for the unknown marginal distribution. More information about the identifiability problem can be found in Tsiatis [34].

2.2 Likelihood function formulation

The theory in this section is selected from Prentice et al. [30] and Chapter 9 in the book by Lawless [23].

Suppose there exist n independent observation units, (T, D) is observed for each unit and right censoring is possible. If a unit is non-censored, the time of failure t_j and the cause of failure d_j is observed. For a right censored observation it is known that the unit survive at least up to time t_j and this time is observed. What happens after t_j is unknown. Let $\delta_j = I(T_j \leq C_j)$, where C_j is the non-negative right censoring time for unit j and I is the indicator function. Under the assumption of independent censoring (see Chapter 2 [23]), the likelihood function can be written

$$L = \prod_{j=1}^{n} f_{d_j}(t_j)^{\delta_j} \bar{F}(t_j)^{1-\delta_j}.$$

From equation (2.2) it follows that

$$L = \prod_{j=1}^{n} \lambda_{d_j}(t_j)^{\delta_j} \bar{F}(t_j)^{\delta_j} \bar{F}(t_j)^{1-\delta_j} = \prod_{j=1}^{n} \lambda_{d_j}(t_j)^{\delta_j} \bar{F}(t_j).$$

Let δ_{jd} indicate if unit j fails due to cause d, $\delta_{jd} = I(D_j = d)$. If subject j is censored or fails due to a competing event, δ_{jd} equals 0. It is not possible that each unit fails due to more than one cause, hence $\delta_j = \sum_{d=1}^{p} \delta_{jd}$. By using equation (2.5) it follows

that

$$\begin{aligned} L &= \prod_{j=1}^{n} \lambda_{d_j}(t_j)^{\sum_{d=1}^{p} \delta_{jd}} \bar{F}(t_j) \\ &= \prod_{j=1}^{n} \prod_{d=1}^{p} \lambda_d(t_j)^{\delta_{jd}} \bar{G}_d^*(t_j) \\ &= \prod_{j=1}^{n} \prod_{d=1}^{p} \frac{g_d^*(t_j)^{\delta_{jd}}}{\bar{G}_i^*(t_j)^{\delta_{jd}}} \bar{G}_d^*(t_j) \\ &= \prod_{d=1}^{p} \Big(\prod_{j=1}^{n} g_d^*(t_j)^{\delta_{jd}} \bar{G}_d^*(t_j)^{1-\delta_{jd}} \Big) \\ &= \prod_{d=1}^{p} L_d, \end{aligned} \tag{2.9}$$

where $g_d^*(t) = \lambda_d(t)\bar{G}_d^*(t)$. From equation (2.9) it is clear that the overall likelihood can be written as a product of p likelihoods, one for each failure cause. The dth likelihood is identical to the standard form of a likelihood with $g_d^*(t)$ as the sub-density function and $\bar{G}_d^*(t)$ as the survival function, although neither $\bar{G}_d^*(t)$ or $g_d^*(t)$ correspond to any observable random variable. The form of the likelihood function is completely specified by the cause specific hazard functions (2.5) and only the hazard functions or functions of them can be estimated directly from the data. Other quantities are non-estimable. If the hazard functions do not depend on the same parameters, the dth likelihood is identical to the likelihood one would obtain by treating failures of other causes than d as censored on their failure time. It is important to notice that no assumptions are required about the interrelation among the failure causes.

2.3 Nonparametric estimation

The theory in this section is selected from Lawless, Chapter 9 [23]. As demonstrated above for parametric estimation, non-parametric estimation of identifiable functions can be done separately for each cause d by treating the other causes as censored.

Kaplan Meier estimate based on the data (t_j, δ_{dj}) can be used to estimate $\bar{G}_d^*(t)$. This is usually not of interest since $\bar{G}_d^*(t)$ is not an observable random variable. However, the result from Kaplan Meier estimate of $\bar{G}_d^*(t)$ can be used to estimate the cumulative cause specific hazard function, Λ_d, see equation (2.5). Another option is to estimate the cumulative cause specific hazard function by a Nelson-Aalen estimator, which takes the form

$$\hat{\Lambda}_d(t) = \sum_{j:t_j \leq t} \frac{\tilde{d}_{jd}}{n_j}, \qquad d \in \{1, ..., p\}, \tag{2.10}$$

where n_j denote the number at risk just before time t_j and $\tilde{d}_{jd}$ is the number of failures of cause d by time t_j.

The variance is typically estimated by

$$\hat{\text{Var}}[\hat{\Lambda}_d(t)] = \sum_{j:t_j \leq t} \frac{\tilde{d}_{jd}}{n_j^2}, \qquad d \in \{1, ..., p\}.$$

The overall survival function can be estimated by the Kaplan Meier estimate on the data (t_j, δ_j), where $0 < t_1 < t_2 < ... < t_n$ are the ordered distinct failure times where failure of any cause occur.

$$\hat{\bar{F}}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{\tilde{d}_j}{n_j}\right), \tag{2.11}$$

where $\tilde{d}_j = \sum_{d=1}^{p} \tilde{d}_{jd}$ is the total number of failures from any cause by time t_j. This estimator is identical to the standard Kaplan-Meier estimator one would obtain treating failures of other causes than d as censored observations.

It follows from equation (2.4), (2.10) and (2.11) that a non-parametric estimate of the cumulative incidence functions is

$$\hat{F}_d(t) = \int_0^t \hat{\bar{F}}(u)\hat{\lambda}_d(u)du = \sum_{j:t_j \leq t} \frac{\tilde{d}_{jd}}{n_j} \hat{\bar{F}}(t_j), \qquad d \in \{1, ..., p\}. \tag{2.12}$$

In 1978 Aalen presented a method to find the variance of the estimate of the cumulative incidence function. Details about the method can be found in Chapter 2 in [29].

Special techniques are required to compare the cumulative incidence functions. Grey, among others, has derived methods to test a covariate in the presence of competing risks. The k-sample test is a non-parametric test that compares weighted averages of the hazard of the subdistribution for the failure of interest. More information can be found in an article by Grey [11].

2.4 Regression models

2.4.1 Cause specific hazard functions

To investigate factors that may affect the risk of failure due to a specific cause in the presence of the competing events, a model analogue to Cox proportional hazards model is desirable.

The result from Section 2.2 implies that standard methods from survival analysis can be used for testing and estimating the cause specific hazard functions, with the modification that failures from other causes than the cause of interest is treated as censored observations. Hence, Cox proportional hazards model can be used to model the effect of covariates on the cause specific hazards functions. With covariates $\mathbf{X} = (x_1, x_2, ..., x_m)$ the model takes the form

$$\lambda_d(t|\mathbf{X}) = \lambda_{d0}(t)e^{\boldsymbol{\beta_d^T}\mathbf{X}}, \qquad d \in \{1, 2, ..., p\}, \tag{2.13}$$

where $\lambda_{d0}(t)$ is the baseline hazard of cause d and $\beta_\mathbf{d}$ is a vector of the coefficients.

Estimation of the regression parameters for cause d, $\beta_\mathbf{d}$, is based on the following partial likelihood approach for ordered untied failure times, $t_1 < t_2 < ... < t_n$,

$$L(\beta_\mathbf{d}) = \prod_{j:d_j\delta_j=d} \frac{e^{\boldsymbol{\beta_d^T}\mathbf{X}_j}}{\sum_{l\in R(v_j)} e^{\boldsymbol{\beta_d^T}\mathbf{X}_l}}. \tag{2.14}$$

$d_j\delta_j$ indicates if subject j fails due to cause d, $\delta_j = I(T_j \leq C_j)$, c_j is the potential censoring time and $v_j = \min(t_j, c_j)$ is the terminal time. The risk set for each time t is defined as $R(t) = \{j : u_j \leq t \leq v_j\}$, where u_j is the left truncation variable for subject j. Briefly, left truncation arises when subjects come under observation only some known time after the natural time origin. More information about left truncation can be found in Chapter 2 in the book by J.P. Lawless [23].

It is worth noting that the results should only be interpreted in terms of the cause specific hazard rates, since the cumulative incidence function for the cause of interest is not a simple function of the cause specific hazard rate for the cause of interest, but also a function of the competing events. This will be described further in Section 2.5.

2.4.2 Cumulative incidence functions

In most cases it is more interesting to assess the effect of covariates on the cumulative incidence function directly. For that reason Fine and Grey [9] introduced a proportional hazard model to the subdistribution hazard function. Conditional on the covariates $\mathbf{X} = (x_1, x_2, ..., x_n)$ the model takes the form

$$w_d(t|\mathbf{X}) = w_{d0}(t)e^{\boldsymbol{\phi_d^T}\mathbf{X}}, \qquad d \in \{1, 2, ..., p\}, \tag{2.15}$$

where $w_{d0}(t)$ is the baseline hazard of the subdistribution of cause d (2.6) and $\phi_\mathbf{d}$ is a vector of the coefficients.

Based on this model, the cumulative incidence functions can be written

$$F_d(t;\mathbf{X}) = 1 - e^{-\int_0^t w_{d0}(u)e^{\boldsymbol{\phi_d^T}\mathbf{X}}du} = 1 - \left[e^{-\int_0^t w_{d0}(t)du}\right]^{e^{\boldsymbol{\phi_d^T}\mathbf{X}}}, \qquad d \in \{1, 2, ..., p\}.$$

It follows that the covariate effect on the cumulative incidence function can be interpreted directly.

Estimation of the model parameters depends on different censoring and truncation scenarios. For estimation, Fine and Grey [9] made a distinction between three types of available data, two of which are not used in practice.

The first scenario is "Complete Data" where all failure times and failure causes are observed. It follows that censoring is absent. Without censoring the partial likelihood function is identical to the partial likelihood in equation (2.14). However, the risk set is augmented by individuals who fail prior to time t by a competing event, with a failure time equal to infinity.

The second scenario is called "Censoring Complete Data" and involves data where censoring is only due to administrative loss-to-follow up. The characteristic of "Censoring Complete Data" data is that the censoring time is known, even for subjects who fail prior to the administrative censoring time. As for the first scenario, the partial likelihood method for the cause specific hazard can be applied with the extended risk set.

The third and "used" scenario is called "Incomplete Data". Incomplete data is data where usual right censoring is present. In Fine and Grey's definition, (2.15), failures of competing events prior to the relevant time stay in the risk set for infinity. For general censoring at random, the time of failure for when a competing event remains in the risk set is not known. For this reason, Fine and Grey [9] proposed weighting by inverse probability of censoring (IPCW) techniques [32] to fit the subdistribution hazard models.

Briefly described, $r_j(t) = I(C_j \geq T_j \wedge t)$ denotes the vital status on individual j at time t, $\wedge$ denotes min. $r_j(t)$ takes the value 1 if it is known that subject j has not been censored or failed prior to time t. $r_j(t)$ takes the value 0 if the status is unknown, i.e. censoring has happened before both T_j and t. Based on this quantity and the Kaplan Meier estimate of the survival distribution of the censoring random variable $\hat{G}(t) = P(C \geq t)$, time dependent weights $\tilde{w}_j$ are defined:

$$\tilde{w}_j = \frac{r_j(t)\hat{G}(t)}{\hat{G}(V_j \wedge t)}, \tag{2.16}$$

where V_j is the minimum of the observed failure time, T_j, and observed censoring time, C_j. The weight is equal to 1 for subjects that have neither failed nor been censored until time t. Subjects that have failed from another cause at time T_j prior to time t get a weight equal to $\frac{\hat{G}(t)}{\hat{G}(T_j)}$. Thus, subjects that have failed due to a competing event prior to time t do not participate fully in the partial likelihood. For

more details see Fine and Gray [9].

Defining these weights gives the following partial likelihood

$$L(\phi_d) = \prod_{j:d_j\delta_d=d} \frac{e^{\boldsymbol{\phi_d^T}\mathbf{X}_j}}{\sum_{i\in R_j^*} \tilde{w}_{ji} e^{\boldsymbol{\phi_d^T}\mathbf{X}_i}}, \tag{2.17}$$

with an extended risk set, R_j^*, that includes subjects that have failed by a competing event by time t.

In 2010, Ronald B. Geskus [10] expanded the likelihood to include left truncation. Geskus defined $\hat{H}(t)$ as the Kaplan Meier estimate of the truncated times by reversing the role of the truncated times and the terminal times. Further, he used the IPCW techniques to derived new weights

$$\tilde{w}_j^* = \begin{cases} 1 & \text{If at risk at } t \\ \frac{\hat{G}(t)\hat{H}(t)}{\hat{G}(V_j\wedge t)\hat{H}(V_j\wedge t)} & \text{If } d \text{ had a competing event observed prior to } t \\ 0 & \text{Otherwise.} \end{cases}$$

It is worth noting that these weights are identical to Fine and Grey's definition (2.16) when left truncation is absent, $\hat{H}(t) \equiv 1$.

2.5 Effect of a covariate

Much of the work of analyzing the effect of a covariate in a competing risks situation have been done by examining the effect of the covariate on the cause specific hazard function, see equation (2.13). Gray [11] noted that a covariate may have different effect on the cumulative incidence function and the cause specific hazard function. The reason for this is that the cumulative incidence function for cause d dependes on the overall survival function, and therefore on the competing events, see equation (2.4).

As an example, assume two competing events and an explanatory variable corresponding to two groups, a and b. The cause specific hazard rates are set to be constant and equal to

$$\lambda_{1a} = \lambda_{2a} = 3, \qquad \lambda_{1b} = 2, \qquad \lambda_{2b} = 1.$$

The first index corresponds to the failure cause and the second index to the group.

The cumulative incidence functions can be calculated directly by equation (2.5)

and (2.4),

$$F_{1a}(t) = \int_0^t \lambda_{1a}(u)\bar{F}_a(u)du = \int_0^t \lambda_{1a}e^{-(\lambda_{1a}+\lambda_{2a})u}du = \int_0^t \lambda_{1a}e^{-(\lambda_{1a}+\lambda_{2a})u}du = \frac{1-e^{-6t}}{2}$$
$$F_{1b}(t) = \int_0^t \lambda_{1b}(u)\bar{F}_b(u)du = \int_0^t \lambda_{1b}e^{-(\lambda_{1b}+\lambda_{2b})u}du = \int_0^t \lambda_{1b}e^{-(\lambda_{1b}+\lambda_{2b})u}du = \frac{2(1-e^{-3t})}{3},$$

and the subdistribution hazard functions by equation (2.6),

$$w_{1a}(t) = -\frac{d(log(1-F_{1a}(t))}{dt} = -\frac{d(\log(\frac{1}{2}+\frac{1}{2}e^{-6t})}{dt} = \frac{6}{1+e^{6t}}$$
$$w_{1b}(t) = -\frac{d(log(1-F_{1b}(t))}{dt} = -\frac{d(\log(\frac{1}{3}+\frac{2}{3}e^{-3t})}{dt} = \frac{6}{2+e^{3t}}.$$

An illustration of the cumulative incidence function for the two groups, cause 1, can be seen in Figure 2.1. The figure shows that the cumulative incidence functions converges towards their marginal probability as t goes to infinity. It can also be seen that the two curves cross after a certain time, $F_{1b}(t) > F_{1a}(t)$ for $t > 0.37$.

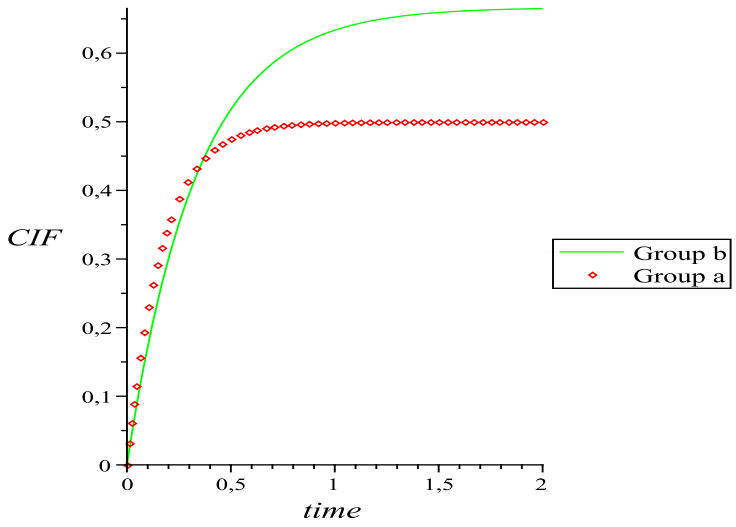


Figure 2.1: Cumulative incidence function for group a and group b, cause 1.

The subdistribution hazard functions can be described entirely in terms of the cumulative incidence functions. Figure 2.2 shows the subdistribution hazard function for the two groups, cause 1. $w_{1b}(t) > w_{1a}(t)$ for $t > 0.16$. The pattern is similar for the cumulative incidence functions and the subdistribution hazard functions. The function for group a is larger for small t while the function for group b is largest for larger t. By comparison $\lambda_{1a} > \lambda_{1b}$ for all t. In other words, the covariate has very

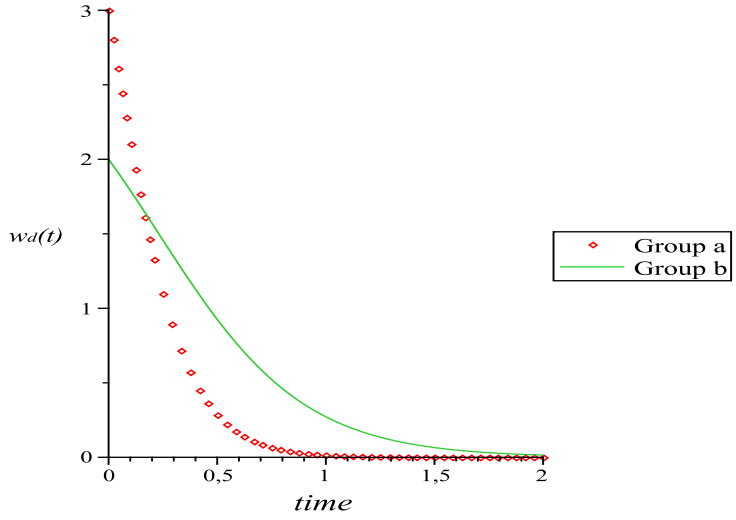


Figure 2.2: Subdistribution hazard function for groups a and group b, cause 1.

different effects on the cause specific hazard function and the cumulative incidence function, and hence the subdistribution hazard function.

The cause specific hazard function is the instantaneous failure rate for a given cause, while the cumulative incidence function is the probability of failure by time t for a given cause. When t goes to infinity a subject has to fail, thus the cumulative incidence function is dependent on the marginal probability for the other failure causes. The cause specific hazard function does not depend on the competing events.

A consequence of different effect from covariates on the cause specific hazard function and the cumulative incidence function is that the null-hypothesis, $H_0 : F_{ia}(t) = F_{ib}(t)$ is not equivalent to $H_0 : \lambda_{ia}(t) = \lambda_{ib}(t)$ unless the survival functions for the groups are equal, $\bar{F}_a(t) = \bar{F}_b(t)$. Proof follows by equation (2.4) and (2.2),

Proof.

$$\lambda_{ia}(t) = \lambda_{ib}(t) \Rightarrow \frac{f_{ia}(t)}{\bar{F}_a(t)} = \frac{f_{ib}(t)}{\bar{F}_b(t)}$$
$$\text{If } \bar{F}_a(t) = \bar{F}_b(t), \text{then}$$
$$f_{ia}(t) = f_{ib}(t) \Rightarrow F_{ia}(t) = F_{ib}(t).$$

$$F_{ia}(t) = F_{ib}(t) \Rightarrow \int_0^t \lambda_{ia}(u)\bar{F}_a(u)du = \int_0^t \lambda_{ib}(u)\bar{F}_b(u)du$$
$$\text{If } \bar{F}_a(t) = \bar{F}_b(t), \text{then}$$
$$\lambda_{ia}(t) = \lambda_{ib}(t)$$

Hence, $H_0 : \lambda_{ia}(t) = \lambda_{ib}(t) \Leftrightarrow F_{ia}(t) = F_{ib}(t)$ if $\bar{F}_a(t) = \bar{F}_b(t)$ □

2.6 Approximate Cox regression

As mentioned in Section 2.4.1, Cox proportional hazard models can be used to model the effect of covariates on the cause specific hazard functions. In the late 90's, Cox regression required large data capacity and were time consuming for large risk sets. For that reason, Krogstad and Lindqvist [15] introduced "Approximate Cox" regression. The Approximate Cox method turns out to be an intuitive method which gives good approximations and is easy to program without "Cox-software". The method also deals with potential time dependency in the regression parameters. For these reasons, the Approximate Cox method will be investigated further on competing risks data. The theory in this section is selected from a lecture by Bo Lindqvist [25].

To estimate the parameters β_d in a proportional hazard model from the observed data, the partial likelihood given in equation (2.14) needs to be maximized. Recall

$$L(\beta_d) = \prod_{j:d_j\delta_j=d} \frac{e^{\beta_{\mathbf{d}}\mathbf{x_j}}}{\sum_{l\in R(v_j)} e^{\beta_{\mathbf{d}}\mathbf{x_l}}} = \prod_{i=1}^{r} \frac{e^{\beta_{\mathbf{d}}\mathbf{x_i}}}{\sum_{l\in R_i} e^{\beta_{\mathbf{d}}\mathbf{x_l}}}. \tag{2.18}$$

In the rightmost equation, the risk set is assumed to be the same for each distinct failure time. The ordered (increasing) failure times of cause d, v_j, are renamed as i = 1,2,...,r.

A good approximation for large risk sets is to replace

$$\frac{1}{N_i}\sum_{l\in R_i} e^{\beta_{\mathbf{d}}\mathbf{x_l}} \approx E[e^{\beta_{\mathbf{d}}\mathbf{X_i}}],$$

where $\mathbf{X_i}$ is a random variable which describes the distribution of the covariates, $\mathbf{X}$, for individuals in the risk set at time t_i, $R_i = \{i : u_i \leq t \leq v_j\}$. N_i is the number at risk by time t_i. A reasonable approximation is to assume the distribution of $\mathbf{X_i}$ as known when the risk set R_i is large. It is important not to confuse $\mathbf{X_i}$ with the observed covariate value for the failures, $\mathbf{x_i}$.

Using the approach, the partial likelihood in equation (2.18) can be written

$$\prod_{i=1}^{r} \frac{e^{\beta_{\mathbf{d}}\mathbf{x_i}}}{N_i \mathrm{E}[e^{\beta_{\mathbf{d}}\mathbf{X_i}}]},$$

and the log-(modified) likelihood

$$\sum_{i=1}^{r} \beta_{\mathbf{d}}\mathbf{x_i} - \sum_{i=1}^{r} \log N_i - \sum_{i=1}^{r} \log \mathrm{E}[e^{\beta_{\mathbf{d}}\mathbf{X_i}}].$$

By differentiating the log-(modified) likelihood with respect to $\beta_{\mathbf{d}}$ and setting it equal to zero, the maximum likelihood equations are obtained,

$$\sum_{i=1}^{r} \left(\mathbf{x_i} - \frac{d}{d\beta_{\mathbf{d}}} \log \mathrm{E}[e^{\beta_{\mathbf{d}}\mathbf{X_i}}] \right) = 0. \tag{2.19}$$

This equation can be solved for β_d when the distribution of the covariates is known and the moment generating function exists. In relation to medical studies, there will often be both categorical and continuous covariates. The normal distribution may be used in the continuous case, while the multinomial distribution is appropriate for categorical variables. As mentioned introductory, the approximate method can reveal potential time dependency in the regression parameters by solving for each parameter locally.

2.6.1 Normally distributed covariates

The covariates are now assumed to be normally distributed with mean μ_i and variance σ_i, $X_i \sim \mathrm{N}(\mu_i, \sigma_i)$, which is reasonable for variables like Body Mass Index or height. It is also assumed that μ_i and σ_i can readily be estimated from the risk set R_i. The moment generating function for the normal distribution is defined as

$$\mathrm{E}[e^{\beta_d X_i}] = e^{(\mu_i \beta_d + \sigma_i^2 \beta_d^2/2)} \Rightarrow \log(\mathrm{E}[e^{\beta_d X_i}]) = \mu_i \beta_d + \frac{\sigma_i^2 \beta_d^2}{2}. \tag{2.20}$$

Inserting equation (2.20) into the maximizing equation, (2.19), and differentiating with respect to β_d gives

$$\sum_{i=1}^{r} \left(x_i - \mu_i - \sigma_i^2 \beta_d \right) = 0.$$

The maximum likelihood estimator, $\hat{\beta}_d$, is obtained by solving for β_d,

$$\hat{\beta}_d = \frac{\sum_{i=1}^{r} (x_i - \mu_i)}{\sum_{i=1}^{r} \sigma_i^2} = \frac{\bar{x} - \bar{\mu}}{\bar{\sigma}^2}. \tag{2.21}$$

Hence, $\hat{\beta}_d$ compares the average covariate value for failures of cause d, $\bar{x}$, with the average covariate value for the population at risk, $\bar{\mu}$, adjusted by the average variance for the population at risk, $\bar{\sigma}^2$.

To make inference on β_d, it is desirable with the distribution of $\hat{\beta}_d$. The following variable is therefore defined for each individual

$$k_{di} = \begin{cases} 1 & \text{If failure of cause } d \text{ occurs at time } i \\ 0 & \text{Otherwise.} \end{cases}$$

It is still assumed that X_i is normally distributed, $X_i \sim \mathrm{N}(\mu_i, \sigma_i)$, and that the basic assumption in Cox regression, $P(k_{di}|x) \propto \exp(\beta_d x)$, applies. It follows that

$$\begin{aligned}
P(x|k_{di}) &\propto P(k_{di}|x)f(x) \\
&\propto \exp(\beta_d x)\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2}) \\
&\propto \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2-2x(\sigma_i^2\beta_d+\mu_i)+(\mu_i+\beta_d\sigma_i^2)^2}{2\sigma_i^2})\exp(\mu_i\beta_d+\beta_d^2\sigma_i^2/2) \\
&\propto \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-(\mu_i+\beta_d\sigma_i^2))^2}{2\sigma_i^2}) \\
&\sim N(\mu_i+\beta_d\sigma_i^2, \sigma_i^2),
\end{aligned}$$

which is the distribution of x for failures of cause d at time i.

Furthermore,

$$\mathrm{E}[\hat{\beta}_d] = \mathrm{E}[\frac{\sum_{i=1}^r (x_i-\mu_i)}{\sum_{i=1}^r \sigma_i^2}] = \frac{\sum_{i=1}^r(\mathrm{E}[x_i]-\mu_i)}{\sum_{i=1}^r\sigma_i^2} = \frac{\sum_{i=1}^r(\mu_i+\beta_d\sigma_i^2-\mu_i)}{\sum_{i=1}^r\sigma_i^2} = \beta_d,$$

and

$$\mathrm{Var}[\hat{\beta}_d] = \mathrm{Var}[\frac{\sum_{i=1}^r (x_i-\mu_i)}{\sum_{i=1}^r \sigma_i^2}] = \frac{1}{(\sum_{i=1}^r\sigma_i^2)^2}\mathrm{Var}[\sum_{i=1}^r(x_i-\mu_i)] = \frac{\sum_{i=1}^r\sigma_i^2}{(\sum_{i=1}^r\sigma_i^2)^2} = \frac{1}{\sum_{i=1}^r\sigma_i^2},$$

assuming independence among the cases. Hence,

$$\hat{\beta}_d \sim N(\beta_d, 1/\sum_{i=1}^r \sigma_i^2).$$

It follows that

$$Z = \hat{\beta}_d\sqrt{\sum_i \sigma_i^2} \sim N(0,1) \qquad \text{under } H_0,$$

can be used to test the null-hypothesis $H_0 : \beta_d = 0$.

In the multivariate normal case, $X_i \sim N_p(\mu_i, \Sigma_i)$, β_d can be estimated similarly as

$$\hat{\beta}_d = \left(\sum_{i=1}^r \Sigma_i\right)^{-1}\sum_{i=1}^r(x_i-\mu_i) = (\bar{\Sigma})^{-1}(\bar{x}-\bar{\mu}).$$

In the same manner as above we get

$$\hat{\beta}_d \sim N_p(\beta_d, (1/r)(\bar{\Sigma})^{-1}),$$

which can be used to make inference on β_d.

2.6.2 Gamma distributed covariates

Assume now that the covariates are gamma distributed, $X_i \sim \text{gamma}(b_i, a_i)$, where b_i and a_i can easily be estimated from the risk set R_i. The moment generating function is given by

$$\mathrm{E}[e^{\beta_d X_i}] = \left(\frac{1}{1 - b_i\beta_d}\right)^{a_i}. \tag{2.22}$$

Inserting the moment generating function, (2.22), into the maximizing equation, (2.19), gives

$$\begin{aligned}\sum_{i=1}^{r} x_i &= \sum_{i=1}^{r} \frac{d}{d\beta_d} \log\left(\frac{1}{1 - b_i\beta_d}\right)^{a_i} \\ &= -\sum_{i=1}^{r} a_i \frac{d}{d\beta_d} \log(1 - b_i\beta_d) \\ &= \sum_{i=1}^{r} \frac{a_i b_i}{1 - b_i\beta_d}.\end{aligned} \tag{2.23}$$

Equation (2.23) must be solved numerically for β_d. However, if $b_i\beta_d$ is small, the approximation $(1 + b_i\beta_d)^{-1} \approx 1 - b_i\beta_d$ may be used. In this case, the problem simplifies to

$$\begin{aligned}&\sum_{i=1}^{r} x_i \approx \sum_{i=1}^{r} a_i b_i (1 + b_i\beta_d) \\ &\qquad \Downarrow \\ &\hat{\beta}_d = \frac{\sum_{i=1}^{r}(x_i - a_i b_i)}{\sum_{i=1}^{r} a_i b_i^2} = \frac{\sum_{i=1}^{r}(x_i - E[x_i])}{\sum_{i=1}^{r} \mathrm{Var}[x_i]},\end{aligned} \tag{2.24}$$

which is a good illustration that equation (2.21) gives a usable result, although it is exactly right only in the normal distribution case.

For small β_d, a Taylor expansion of the moment generating function around $\beta_d = 0$ gives the same result. This will be described further in Section 2.6.5. As for the normal distribution, β_d will be a comparison of the observed covariate value for failure of cause d, with the average covariate value for the population at risk. The exponential and chi-squared distribution are simplifications of the gamma distribution, hence β_d can easily be estimated from simplifications of equation (2.23).

2.6.3 Binomially distributed covariates

It is now assumed that X_i is a categorical variable with 2 categories. It is assumed within the population that

$$P(X_i = 1) = 1 - P(X_i = 0) = p_i,$$

and the observed x_i is either 0 or 1, i.e. success or not.

Using the moment generating function for the binomial distribution we get

$$\mathrm{E}[\exp(\beta_d x_i)] = p_i e^{\beta_d} + 1 - p_i. \tag{2.25}$$

Inserting equation (2.25) into equation (2.19) gives

$$\sum_{i=1}^{r} \left(x_i - \frac{d}{d\beta_d} \log\left(p_i e^{\beta_d} + 1 - p_i\right)\right) = 0$$
$$\sum_{i=1}^{r} x_i = \sum_{i=1}^{r} \frac{p_i e^{\beta_d}}{p_i e^{\beta_d} + 1 - p_i},$$

which has to be solved numerically for β_d.

With $p_i = p$ for all i, this simplifies to

$$\sum_{i=1}^{r} x_i = \sum_{i=1}^{r} \frac{p e^{\beta_d}}{p e^{\beta_d} + 1 - p}$$
$$\bar{x} = \frac{p e^{\beta_d}}{p e^{\beta_d} + 1 - p}$$
$$\Downarrow$$
$$\beta_d = \log\left(\frac{\bar{x}(1-p)}{p(1-\bar{x})}\right) = \log\left(\frac{\bar{x}}{1-\bar{x}}\right) - \log\left(\frac{p}{1-p}\right).$$

In general, a Taylor expansion of the moment generating function around $\beta_d = 0$ for small β_d gives

$$\beta_d \approx \frac{\sum_{i=1}^{r}(x_i - E[x_i])}{\sum_{i=1}^{r} \mathrm{Var}[x_i]} = \frac{\sum_{i=1}^{r}(x_i - p_i)}{\sum_{i=1}^{r} p_i(1-p_i)}, \tag{2.26}$$

which is a comparison between the observed failure cases and the related risk set. Further description is found in Section 2.6.5.

2.6.4 Multinomially distributed covariates

It is now assumed that X_i is a categorical variable with k possible outcomes, i.e. an extension of the binomial distribution. X_i is a random variable that indicates which outcome occurred, represented by $X = (X_1, X_2, ..., X_{k-1})$ where $X_j = 1$ and $X_s = 0$ $(j \neq s)$ for the j'th category (j = 1,2,..., $k-1$). For the reference category, k, $X_s = 0 \forall s$. Furthermore, $\mathrm{P}(x_j = 1, x_s = 0 \forall s \neq j) = p_j$, j = 1,2,...,$k-1$ and $\mathrm{P}(x_j = 0 \forall j) = 1$ - $\sum_{j=1}^{k-1} p_j$. The maximizing problem, (2.19), changes to

$$\sum_{i=1}^{r} x_i = \sum_{i=1}^{r} \frac{d}{d\beta_{dj}} \log \mathrm{E}[e^{\sum_{j=1}^{k-1} \beta_{dj} X_{ij}}], \qquad j = 1, ..., k-1, d = 1, ..., p,$$

which, by the use of the moment generating function gives

$$\begin{aligned}\sum_{i=1}^{r} x_i &= \sum_{i=1}^{r} \frac{d}{d\beta_{dj}} \log\Big(\sum_{j=1}^{k-1} p_j e^{\beta_{dj}} + p_k\Big) \\ &= \sum_{i=1}^{r} \frac{p_j e^{\beta_{dj}}}{\sum_{j=1}^{k-1} p_j e^{\beta_{dj}} + p_k} \\ &= \sum_{i=1}^{r} \frac{p_j e^{\beta_{dj}}}{\sum_{j=1}^{k-1} p_j e^{\beta_{dj}} + 1 - \sum_{j=1}^{k-1} p_j} \qquad j = 1, ..., k-1, d = 1, ..., p. \end{aligned} \tag{2.27}$$

Equation (2.27) must be solved numerically for the estimated coefficients in each distinct group, $\beta_{dj}, j = 1, ..., k-1$. $\beta_{dk} = 0$ and corresponds to the reference category.

2.6.5 General distribution for the covariates

In principle, equation (2.19) can be solved for all distributions where the moment generating function exists.

The moment generating function of a random variable X is defined as

$$M_X(t) := \mathrm{E}\left[e^{tX}\right], \quad t \in \mathbb{R},$$

wherever this expectation exists. The logarithm of the moment generating function, $G(t) = \log(M_X(t)) = \log(\mathrm{E}[e^{tX}])$, also known as the cumulant generating function, have many useful properties. For example, the first moment is the mean, $G'(0) = \mathrm{E}[X]$, and the second moment is the variance, $G''(0) = \mathrm{Var}[X]$. The third moment indicates the skewness of the distribution, $G'''(0) = \mathrm{E}[X - \mathrm{E}[X]]^3$.

Taylor expansion around 0 of $G(t)$ and using $t = \beta_d$ gives

$$\begin{aligned}\log \mathrm{E}[e^{\beta_d X}] = G(\beta_d) &\approx G(0) + G'(0)\beta_d + \frac{G''(0)\beta_d^2}{2} + o(\beta_d^3) \\ &= \mu\beta_d + \sigma^2\beta_d^2/2 + o(\beta_d^3).\end{aligned}$$

This can be used directly in equation (2.19). For small β_d it follows that

$$\hat{\beta}_d \approx \frac{\sum_{i=1}^{r}(x_i - \mu_i)}{\sum_{i=1}^{r}\sigma_i^2},$$

where μ_i and σ_i^2 are the mean and variance, respectively, of the distribution of X_i. These can in practice be estimated unbiased from the risk sets by the mean and the empirical variance. It is assumed that $o(\beta_d^3)$ is negligible, i.e. that the effect of skewness can be ignored. Clearly, the Approximate Cox method for normal distributed covariates seems to be valid regardless of the distribution of the covariates.

2.6.6 Local estimation of β_d

A local estimation of β_d for each failure may suggest if the parameter is time dependent. For a univariate normal distributed covariate, β_d is a weighted average of variables R_i,

$$R_i = \frac{x_i - \mu_i}{\sigma_i^2}. \tag{2.28}$$

If the model is correct, R_i is normally distributed, $N(\beta, 1/\sigma_i^2)$, and should be following a straight horizontal line at height β_d. If this is not the case, β_d is time dependent and should be modeled as a function of t.

2.7 Approximate Fine and Grey regression

As mentioned in Section 2.4.2, a proportional hazard model corresponding to the subdistribution hazard can be used to model the effect of covariates on the cumulative incidence functions. This section provides approximate analysis for the subdistribution hazard functions. A discrete version of the subdistribution hazard rate corresponding to equation (2.6) is

$$\begin{aligned} w_d(t;x) &= P(T = t, D = d | T \geq t \cup (T < t \cap D \neq d), X = x) \\ &= w_{d0}(t) e^{\gamma_d x}. \end{aligned}$$

The cumulative incidence functions are related to the discrete subdistribution hazard functions as follows

$$\begin{aligned} &P(T \leq t \cap D = d | X = x) = 1 - P(T \geq t+1 \cup D \neq d | X = x) \\ &= 1 - P((T \geq t+1 \cap T \geq t \cap T \geq t-1 \cap ... \cap T \geq 1) \cup D \neq d | X = x) \\ &= 1 - \prod_{s=0}^{t} P(T \geq s+1 \cup D \neq d | (T \geq s \cup D \neq d) \cap X = x) \\ &= 1 - \prod_{s=0}^{t} [1 - P(T \leq s \cap D = d | (T \geq s \cup D \neq d) \cap X = x)] \\ &= 1 - \prod_{s=0}^{t} [1 - P(T = s \cap D = d | (T \geq s \cup D \neq d) \cap X = x)] \\ &= 1 - \prod_{s=0}^{t} [1 - P(T = s \cap D = d | (T \geq s \cup (T < s \cap D \neq d)) \cap X = x)] \\ &= 1 - \prod_{s=0}^{t} (1 - w_d(s;x)). \end{aligned}$$

The form of these functions is similar to the continuous cumulative incidence functions described in equation (2.7).

In the Approximate Cox method, the densities $f(x|T \geq t)$ are considered known. Due to different risk set, the approximate analysis of the subdistribution hazard considers the densities $f(x|T \geq t \cup (T < t \cap D \neq d))$ as known. By using the following relation for disjoint events, A and B,

$$P(C|A \cup B) = \frac{P(C \cap (A \cup B))}{P(A \cup B)} = \frac{P(C \cap A) + P(C \cap B)}{P(A \cup B)} = \frac{P(C|A)P(A) + P(C|B)P(B)}{P(A \cup B)},$$

(which can easily be generalized to more than two disjoint events in the union) we get

$$\begin{aligned} f(x|T \geq t \cup (T < t \cap D \neq d)) &= f(x|(T \geq t) \cup_{s=1}^{t-1} (T = s \cap D \neq d)) \\ &\propto f(x|T \geq t)P(T \geq t) + \sum_{s=1}^{t-1} f(x|T = s \cap D \neq d)P(T = s \cap D \neq d). \end{aligned}$$

In this case disjoint means that the event of surviving up to time t is disjoint from the event of failing due to a competing event prior to time t, which is obviously true.

The probability of failing at time t of a competing event, $P(T = s \cap D \neq d)$, is not directly estimable from our data. However, using the following extension it is estimable

$$\begin{aligned} P(T = s \cap D \neq d) &= P(T = s \cap T \geq s \cap D \neq d) \\ &= P(T = s \cap D \neq d|T \geq s)P(T \geq s). \end{aligned}$$

It follows that

$$\begin{aligned} &f(x|T \geq t \cup (T < t \cap D \neq d)) \propto \\ &f(x|T \geq t)P(T \geq t) + \sum_{s=1}^{t-1} f(x|T = s \cap D \neq d)P(T = s \cap D \neq d|T \geq s)P(t \geq s) \\ &\propto f(x|T \geq t) + \sum_{s=1}^{t-1} f(x|T = s \cap D \neq d)P(T = s \cap D \neq d|T \geq s)\frac{P(T \geq s)}{P(T \geq t)} \\ &\propto f(x|T \geq t) + \sum_{s=1}^{t-1} f(x|T = s \cap D \neq d)\frac{P(T = s \cap D \neq d|T \geq s)}{\prod_{i=s}^{t-1}(1 - P(T = i|T \geq i))} \\ &\propto f(x|T \geq t) + \sum_{s=1}^{t-1} f(x|T = s \cap D \neq d)w_s, \end{aligned} \tag{2.29}$$

where

$$w_s = \frac{P(T = s \cap D \neq d | T \geq s)}{\prod_{i=s}^{t-1}(1 - P(T = i | T \geq i))}.$$

All the ingredients are now estimable from the data. The approximate subdistribution hazards can be computed by equation (2.19), with modified means and variances as follows

$$\frac{\sum_{i=1}^{N_t} g(x_i) + \sum_{s=1}^{t-1}\left[\sum_{i=1}^{n_s^{\neq d}} g(x_{is}, D \neq d)\frac{N_t}{N_s}\frac{K(s)}{K(t)}\right]}{N_t + \sum_{i=s}^{t-1}\frac{N_t}{N_s}\frac{K(s)}{K(t)}n_s^{\neq d}}. \tag{2.30}$$

In equation 2.30, $K(t)$ is the Kaplan Meier estimator of P$(T \geq t)$: $K(t) = \prod_{i=1}^{r}(1 - \frac{n_i}{N_i})$, n_i the number of failures by time i, $N_i = \#R_i$ the number at risk by time i and $n_i^{\neq d}$ the number of failures of other causes than d by time i. The mean and variance is calculated by taking the appropriate formula for $g(x_i)$. The estimation equation,2.30, is adjusted such that the probability density function sums to 1.

It is worth noting that Approximate Cox regression consider $f(x|T \geq t)$ as known due to the size of the risk set. Now, the terms $f(x|T = s \cap D \neq d)$ are added. These terms are not necessarily known with the same accuracy as $f(x|T \geq t)$, hence some approximation error may occur.

An alternative formulation of the partial likelihood function for the subdistribution hazards event type d follows from equation (2.30)

$$\prod_{j:d_j\delta_j=d} \frac{e^{\gamma_d x_j}}{\sum_{l\in R(v_j)} e^{\gamma_d x_l} + \sum_{l:v_l<v_j;d_j\delta_j\neq d,0} \frac{N(v_j)}{N(v_l)}\frac{K(v_l)}{K(v_j)}e^{\gamma_d x_l}}.$$

Subjects that have not failed to any event or been censored prior to time t gets a weight equal to 1. Subjects that have already failed due to a competing event prior to time t are weighted $\frac{N(v_j)}{N(v_l)}\frac{K(v_l)}{K(v_j)}$.

The estimation of $f(x|T \geq t)$ and $f(x|T = s \cap D \neq d)$ are done by using the available data at time t and s, respectively. This means that there is an underlying assumption that $f(x|T \geq t) = f(x|T \geq t, L \leq t \leq C)$, where L, C are the left truncation and censoring time, respectively. To have $f(x|T \geq t) = f(x|T \geq t, L \leq t \leq C)$ it must be assumed, for example, that T and (L, C) are conditionally independent given X, and that (L, C) is independent of X. It is also needed that $f(x|T = s \cap D \neq d)$ $= f(x|T = s \cap D \neq d \cap L \leq s \leq C)$ which requires, for example, that (T, D) is conditionally independent of (L, C) given X, and that (L, C) is independent of X.

More details and another (weaker) sufficient condition for $f(x|T \geq t) = f(x|T \geq t, L \leq t \leq C)$ can be found in [25].

Chapter 3

Data description

This chapter contains information about the data set analyzed in this study and a review of the various explanatory variables. The information is selected from the doctoral thesis of Gunnar Kvåle [16].

3.1 Material

In 1955, The Norwegian Cancer Society [1] decided to conduct a screening program for early diagnosis of breast cancer, henceforth HUNT0. All Norwegian women aged 20-69 years by the first of January 1956 in the counties of Vestfold, Vest-Agder, Aust-Agder and Nord-Trøndelag were invited to participate.

In the years 1956 to 1959 the participants were interviewed according to a standard questionnaire concerning reproductive factors and demographic data. They also had a clinical breasts examination carried out by a physician. The population in Nord-Trøndelag was offered three screening examinations, Aust-Agder two and Vestfold one. In the counties where multiple examinations were offered, missing and updated information was added from the last examination(s).

The attendance in Vest-Agder was only 51.7 % and the screening program was originally organized as a pilot project, it was therefore decided to exclude all participants from this county. Due to incomplete data on reproductive factors, the survey is confined to women older than 27 years by the first of January 1956.

In 1964 the official registration number was introduced in Norway. This number was retrieved for 92573 women in the three relevant counties by the start of follow up, first of January 1961. At this time 85063 women were still alive. Among them, 63090 filled out the questionnaire. This yields a response rate of 74.2 %. Women who did not fill out the questionnaire are taken out of the analysis due to missing information.

By using the official personal registration number, cancer occurrences were linked to the survey from the Cancer Registry of Norway. The data set was also linked to the Central Bureau of Statistics for information regarding date of deaths or emigration. Data on height and weight was assigned to the study from a study organized by the National Mass Radiography Service in 1963-1975.

The latest cancer diagnosis was recorded in February 2010 and this is therefore set to the end date for the follow-up period. This means that the follow-up period extended from January 1 1961 to February 15 2010, a period of almost 50 years. Breast, ovarian and uterine cancer are the selected competing events and only cancer occurrences diagnosed in the follow-up period are considered. Thus, women with cancer in the uterine, ovarian or breast before the start of follow up are removed from the data set. Women with other types of cancer before 01.01.1961 are still in the risk set.

Participants that reported a missing uterine or ovaries are eliminated from the analysis in order to get the same risk set for all competing risks. This leaves 61457 women for analysis.

Women reported dead or emigrated are right censored on the terminal date. Women alive and without a breast, uterine or ovarian cancer diagnosis at the end of survey are right censored on the end of follow-up. Overall, 56273 women are right censored, this is approximately 92 percent of the participants.

The data set have been checked for inconsistent information by the use of cross tabulation. If the inconsistent information was not obvious, the variable was regarded as missing.

3.2 Study variables

3.2.1 Age

Age will be used as the time parameter in this analysis because previous studies have shown that the risk of cancer depends on age. A histogram of age by the first of January 1961 for women participating in the study can be seen in Figure 3.1. The youngest woman was 32 years by the start of follow up, while the oldest woman was 74 years. The average age was 49 years.

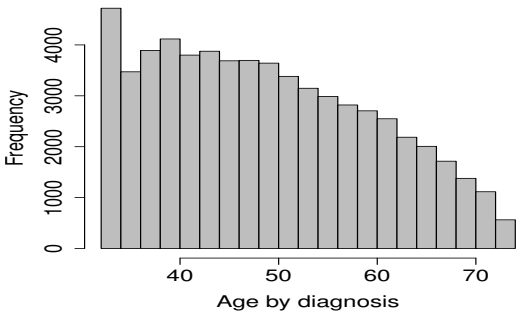


Figure 3.1: Age of participants by the first of january, 1961.

3.2.2 Reproductive variables

Age at first and last birth

Figure 3.2 and 3.3 shows histogram of age at first and last birth, respectively. The average age of first birth is 26.27 years, while the average age of last birth is 32.45 years. For women with one child, age of first and last birth is identical, this applies to 10326 women. It is reasonable to assume that a combination of age at first birth and parity is correlated to age at last birth. There are 14632 missing values for both age at last and first birth, this seems likely since 11063 women are registered as childless. Age at last birth relates to the last child born before the woman attended the screening program. For premenopausal women this will not necessarily coincide with their actual last birth.

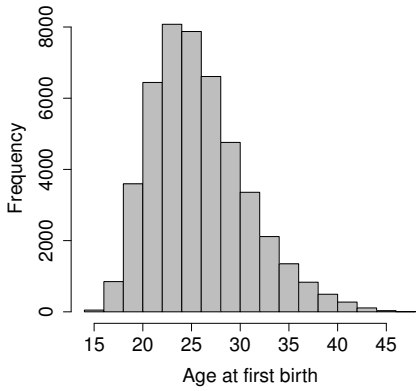


Figure 3.2: Histogram of age at first birth.

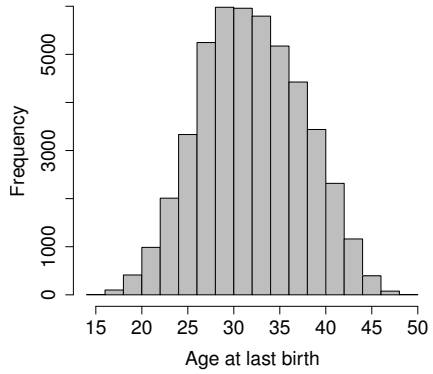


Figure 3.3: Histogram of age at last birth.

Table 3.1 shows that high parity is associated with early first birth and late last birth. Low parity is associated with late first birth and early last birth. Obviously, late age at first birth is associated with late age at last birth.

Age at first birth	Mean parity (SD)	Age at last birth	Mean parity (SD)
$<$ 20 yr	3.62 (2.04)	$<$ 20 yr	1.08 (0.28)
20-24 yr	3.09 (1.70)	20-24 yr	1.53 (0.71)
25-30 yr	2.47 (1.34)	25-30 yr	2.07 (0.99)
30-34 yr	2.08 (1.10)	30-34 yr	2.61 (1.27)
35-40 yr	1.62(0.85)	35-40 yr	3.25 (1.67)
40 + yr	1.20 (0.48)	40 + yr	4.09 (2.29)

Table 3.1: Brief overview of data by reproductive variables.

Parity
Parity refers to the number of times a woman has given birth. A multiple pregnancy counts the same as a pregnancy with one embryo. A nulliparous woman is a woman who has never completed a pregnancy beyond 20 weeks, while a multiparous woman is a woman with more than one pregnancy completed. In this dataset, the average parity is 2.25. A histogram of parity can be seen in Figure 3.4. There are missing information about parity for 1272 women.

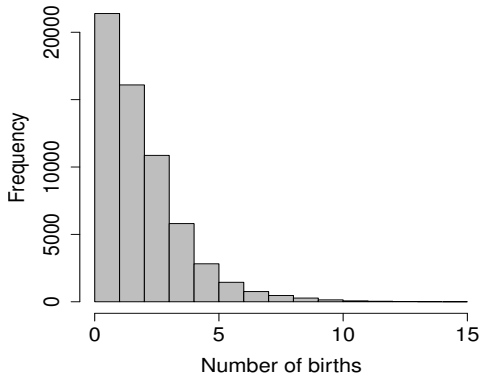


Figure 3.4: Histogram of parity by the first of january, 1961.

3.2.3 Demographic variables

Marital status
Marital status is divided into 10 different subgroups, never married, married, widow, separated and various combinations of these variables. The number of observations in each group are very unevenly distributed, it is for example 49734 observations in one group while another has 8 observations. For this reason, convergence problems occur and further grouping is needed. The new grouping of marital status will be; never married, married and separated / widow. 264 women did not answer the question about marital status.

Residence
Nord-Trøndelag, Vestfold and Vest-Agder are three of 19 Norwegian counties. Figure 3.5 shows their location. It is reasonable to assume that there will be differences in disease prevalence for different counties due to inheritance and kinship. Table 3.2 shows a brief summary of the variable county. The information indicates that the risk of cancer is approximately the same in each of the three relevant counties.

Figure 3.5: Study areas.

	Nord-Trøndelag	Aust-Agder	Vestfold
Participants	21142	13418	26897
Breast cancer	1201	715	1491
Uterine cancer	323	204	407
Ovarian cancer	266	198	379
Fraction cancer	8.5 %	8.3 %	8.5 %

Table 3.2: Brief overview of data by residence.

Residence is also recorded as either urban or rural place, 12956 with urban residence and 48501 with rural. Neither Nord-Trøndelag, Vestfold nor Vest-Agder has any big cities, it is therefore reasonable to assume that the differences between the population in urban and rural residence is small.

Occupational socioeconomic status
Occupation is recorded as either own or husband's occupation and divided into 7 different subgroups; professional/private enterprise, clerical work, fishing/ship officer/cress, farm & forestry work, industrial work, domestic & other work and not specified work. There seems to be about as many women in each occupational group.

3.2.4 Height and weight

Information regarding height and weight were assigned to the study in the years 1963 - 1975 by a health survey organized by the Mass Radiography service. Many of the women did not participate in this survey, hence there are a lot of missing values when it comes to height and weight, more precisely 12481.

At the examination, height was measured in the nearest centimeter and weight was measured to the nearest kilogram. 1693 women are registered with some sort of "disabilities" such as lame, pregnant, bent neck, back or knees. These disabilities are natural biases in a population and will not be classified as errors, the women will therefore not be removed from the analysis. It is also reasonable to assume that the weight will vary during a longer period of time, and one may therefore get some faults in the analysis.

Body Mass Index (BMI) is used to estimate a healthy body weight relative to a person's height. It is defined as weight in kilograms divided by the square value of height in meters (kg/m^2) and is commonly used to classify under-weight, overweight and obesity in adults.

The World Health organization [3], classifies BMI into under-weight ($< 18.5\ kg/m^2$), normal range ($18.5\ kg/m^2$ - $24.99\ kg/m^2$), overweight($\geq 25\ kg/m^2$) and obese (≥ 30

kg/m^2). This classification will partly be used in the analysis, with the exception that under-weight and normal range are further grouped into lean ($< 25\ kg/m^2$).

A histogram of BMI for the participants is given in Figure 3.6. The average BMI for the participants is 26.47 kg/m^2. Table 3.3 summarize parts of the information available on BMI. It is clear from the table that high body mass index is related to early menarche, late menopause and high parity. Comparing fraction of obese women with fraction of women with ovarian cancer indicates that ovarian cancer is associated with low BMI. By the same comparison it seems like the risk of uterine cancer is associated with obesity.

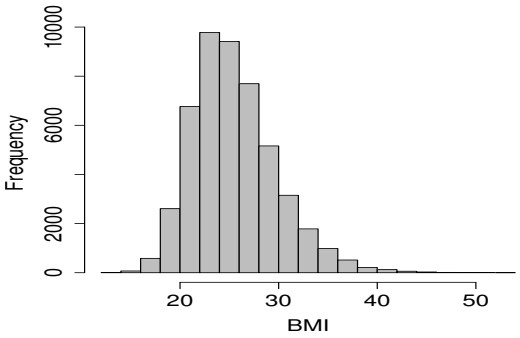


Figure 3.6: Histogram of BMI.

BMI	Lean	Overweight	Obese
Occurrence	19812	19906	9258
Fraction, %	40.5	40.6	18.9
Mean BMI, kg/m^2 (SD)	22.15(1.68)	26.74(1.37)	32.71 (2.97)
Given breast cancer, %	38.2	41	20.8
Given uterine cancer, %	32.4	40.2	27.4
Given ovarian cancer, %	44.2	40.1	14.8
Mean number of births (SD)	1.95 (1.55)	2.30 (1.76)	2.64 (1.99)
Mean age menarche, yr (SD)	14.34(1.40)	14.17 (1.39)	13.98 (1.40)
Mean age menopause, yr (SD)	47.96 (4.37)	48.31 (4.21)	48.34 (4.30)

Table 3.3: Summary of available information on BMI.

3.2.5 Reproductive period

The reproductive period is the period between menarche and menopause. In HUNT0, there exist information about the age of menarche for most of the women, except 1678. The average age for menarche was 14.22 years. Figure 3.7 shows a histogram of age at menarche. Many of the women are too young to have been through menopause by the examination date, it follows many missing values, more precisely 40398. Consequently, information on some of the reproductive variables may be

incomplete, as for example age at last birth and number of births. The average age at menopause in this cohort is 48.26 years. A histogram of age at menopause can be seen in Figure 3.8.

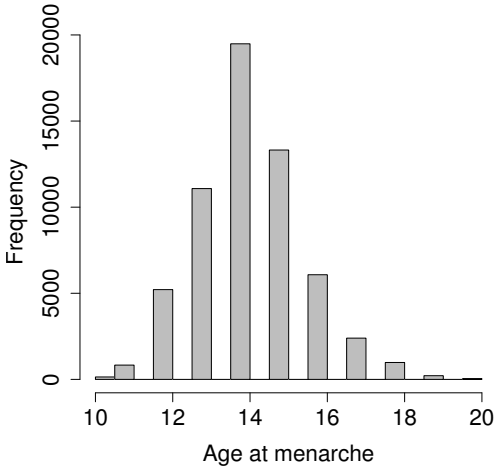


Figure 3.7: Histogram of age at menarche

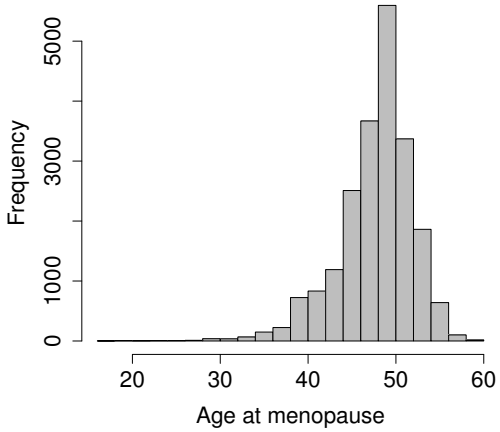


Figure 3.8: Histogram of age at menopause.

There exist information on "duration from menarche to first birth" and "duration from last birth to menopause", these variables are however found to be strongly correlated to "age at first birth" and "age at menopause", respectively. Hence, they will not be studied further in this analysis.

3.2.6 Lactation

Lactation is collected as the length of lactation (in months) for the first, second and third child. The mean duration of lactation is also collected. 4160 women have not breast-fed, this is approximately 9.5 % of parous women. The average length of lactation is 5.85 months. There is no information regarding lactation for 16654 women, this can be explained by the fact that 11063 women are childless.

3.2.7 Abortion

HUNT0 contains information of number of abortions. The explanatory variable *Abortion* is divided into 10 subgroups, 0 to 9 abortions. It is not distinguished between spontaneous abortion or induced abortion. In order to avoid convergence problems, *Abortion* will be regrouped into "abortion" or "no abortion". 11846 women have carried out an abortion, while 47268 have not. There is missing information about abortion for 2343 women.

Chapter 4

Competing risks

HUNT0 provides ideal material for competing risks analysis because of thorough follow-up of cancer occurrences by the Cancer Registry of Norway [1], and almost 85 percent registered deaths. Considering every cancer cause as a competing event will be complicated and meaningless due to lack of observations is each group. Three competing risks with similar risk factors are chosen for this analysis, they will be presented in the following. Information regarding cancer occurrences, incidence rates and prevalence in Norway is selected from the Cancer Registry of Norway [1].

4.1 Breast cancer

4.1.1 General information about breast cancer

The World Health Organization, WHO [3] states that breast cancer is the most common cancer in women worldwide and it comprises 16 % of all female cancers. In Norway, there were 2763 new cases in 2008, which is approximately 23 % of all female cancer diagnosis in 2008. This makes it the most common female cancer in Norway. The cumulative risk of developing cancer by age 75 in Norway is 8.1 %. Men can also get breast cancer, though it is a rare case. 21 Norwegian men got a breast cancer diagnosis in 2008.

Figure 4.1 shows the average number of new breast cancer cases and the average age specific incidence rate per 100 000 person years in Norway from 2004 to 2008. The data is divided into age groups of five years. The incidence rate is defined as the number of new cases of a disease in a population within a defined time period, it indicates the risk of the disease. As the figure shows, the risk increases with age and approximately 80 % of women affected are over 50 years. The risk is highest in the age group 65-69 years. The risk is very low for women younger than 30 years, it comprises only 0.4 % of registered breast cancer cases. Most breast cancers cases occur in the years around menopause, 50-65 years.

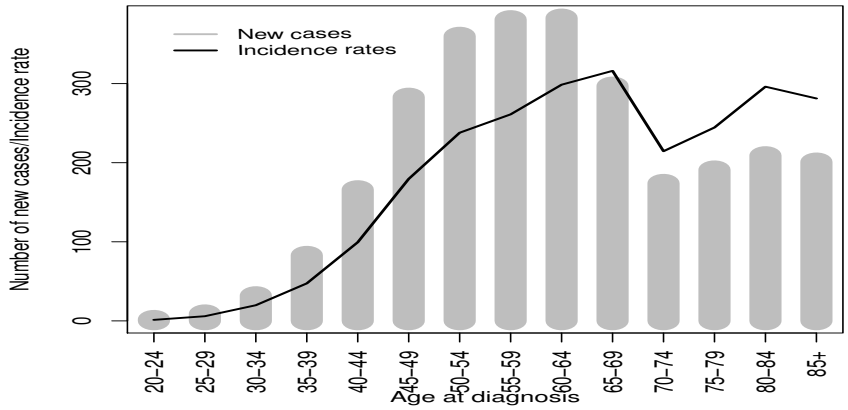


Figure 4.1: Average number of new breast cancer cases and age specific incidence rate per 100 000 person years in Norway from 2004 to 2008.

Figure 4.2 shows the incidence rates of breast cancer per 100 000 person years from 1954 to 2008, divided into 5 year periods. It is obvious from the figure that the number of breast cancer cases in Norway has increased rapidly the recent years. The incidence rate have doubled (from 36.6 to 74.7), and the number of cases have tripled from a average of 868 new cases in 1954-1958 to an average of 2753 new cases in 2004-2008.

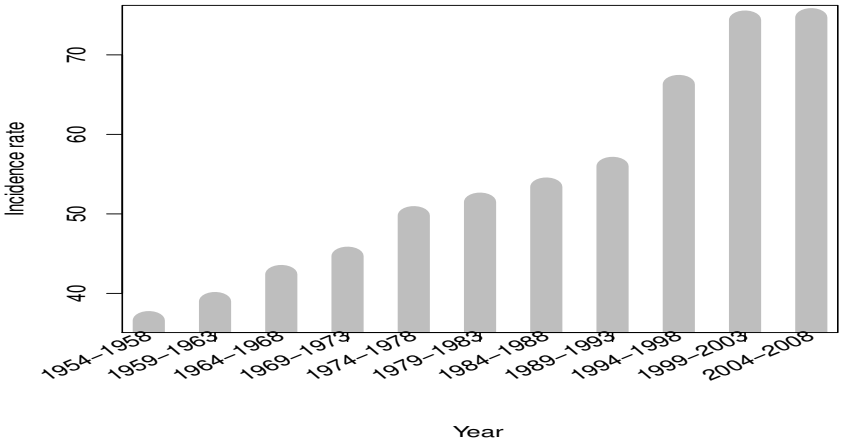


Figure 4.2: Average incidence rate per 100 000 person years, breast cancer, 1954 to 2008.

The probability of surviving breast cancer has also changed in recent years. The probability of surviving breast cancer in the period 1969 to 1973 was 65.2 % in average, while it was 87.8 % in the period 2004 to 2008. A large part of the increase may be due to the introduction of mammography screening in the early 2000's. Earlier diagnosis gives better prognosis.

Prevalence data relates to the number of people in a defined population alive at a specific time with a specific diagnosis. In Norway, 34890 women with a breast cancer diagnosis were alive on the 31.12.2008. Out of these women, 13674 had lived with the diagnosis for more than 10 years.

4.1.2 Current knowledge

Despite an enormous research effort, the epidemiology of breast cancer is only partially understood. There is however some well known risk factors to be aware of.

The risk of breast cancer is strongly related to age. It is a doubling in risk for every 10th year until the menopause, thereafter the rate of increase slows dramatically, [27].

The incidence of breast cancer differs strongly among countries. The incidence rate may be up to five times higher in developed countries compared to less developed countries. It has been recorded increased occurrence of breast cancer for people who move from an area of low incidence to areas with high incidence, hence the environmental factors are of greater importance than the genetic factors, [27].

Much research attention is focused on reproductive factors and many researchers have found that the risk of breast cancer decreases among multiparous women compared to nulliparous women, [20, 31]. Age at first birth has long been considered the major reproductive risk factor in breast cancer, with increasing risk for increasing age at first birth, [27, 31]. Kvåle and Heuch [17] observed that there is no initial significant association between risk of breast cancer and age at last birth. There is however a significant reduced risk if adjusted for parity, the relationship is described as very complex.

A number of studies have concluded that early menarche and late menopause (long reproductive period) increase the risk of breast cancer, [19, 4].

It has been proven that there is accumulation of breast cancer in some families and that up to 10 % of breast cancer cases in Western countries are related to inheritance. A family history of breast cancer in a first degree relative (mother, sister or daughter) before the age of 50 has been associated with approximately a doubling of risk, [27].

Several studies have shown that obesity is associated with increasing risk of breast cancer in postmenopausal women, and decreasing risk for premenopausal women, [27, 4].

Some studies have found that there is a slight reduction in risk of breast cancer for breast-feeding women, [4]. Kvåle and Heuch [18] reached the same result, but their overall impression, based on data, was that breast feeding is not strongly related to the risks of breast cancer.

Other risk factors have also received attention among researchers, as for example alcohol intake, healthy diet, oral contraceptive use, hormone replacement therapy, birth weight, previous benign breast disease, radiation and lack of exercise. These factors not are studied further in this analysis.

4.1.3 Breast cancer in HUNT0

During 48 years of follow up, approximately 5.55 % of the population (3407 women) was diagnosed with breast cancer. A histogram of number of breast cancer cases for each age is shown in Figure 4.3. Compared to the age distribution in Figure 4.1, it is obvious that the average age of breast cancer is displaced.

The Cancer Registry of Norway [1] states that the average age at breast cancer diagnosis in Norway in 2008 was 61.5 years. The average age in HUNT0 is 69.5 years. The difference can be explained by several factors. Information from the Cancer Registry comes from an open population, i.e. the entire population with the "flow" of people who are born and die. HUNT0 is a closed population, it is limited to women born between 1886 and 1928 who were alive in 1960. This means that all women who developed breast cancer before 1960 or died of the disease before 1960 are not included. This will, because of the age distribution, almost exclude all women with a diagnosis before the age of 70 years, which will affect the average age considerably.

Another factor that might affect the average age is the introduction of breast screening with mammography of women aged 50-69 years in Norway in the early 2000's. The women in HUNT0 have not been screened because they are too old. Screening leads to earlier detection of tumors, which also lower the average age at diagnosis.

Hormone supplements during menopause increase the risk of breast cancer and the growth rate of tumors. During the last 20-30 years, hormone supplement has become more common in Norway. Tumors that grow faster are detected earlier and a lower average age follows.

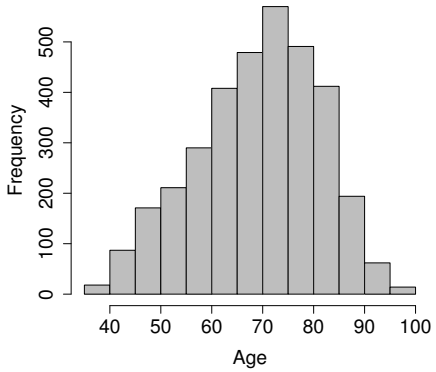


Figure 4.3: Histogram of age by breast cancer diagnosis in HUNT0.

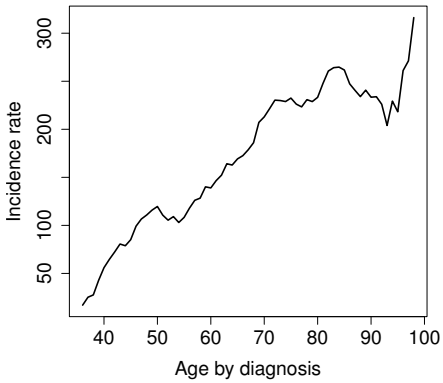


Figure 4.4: Incidence rate of breast cancer per 100 000 person year.

Figure 4.4 shows the incidence rates of breast cancer for different ages, the shape is approximately similar to the incidence curve in Figure 4.1, with the exception that the peak point is shifted, as explained above. The plot is smoothed by taking the average of the two nearest ages on both sides.

4.2 Cancer of the Uterus

4.2.1 General information about uterine cancer

Cancer of the corpus uteri is the fourth most common cancer in women in Norway with 716 new cases diagnosed in 2008, accounting for approximately 6% of all female cancers. The cumulative risk of developing uterine cancer by age 75 for Norwegian women is 2.1 %. Cancer research UK [4] states that uterine cancer is primarily a cancer of the developed world with incidence rates double those of the less developed countries.

Figure 4.5 shows the average number of new uterine cancer cases and the average age specific incidence rates in Norway from 2004 to 2008 divided into age groups of five years. From the figure it can be seen that the risk of uterine cancer increases with increasing age. The majority of cases are diagnosed in women older than 50 years (92.5 % of the cases). Women younger than 50 years are relatively rarely affected by the disease. The incidence rate decreases after the age of 70-74 years.

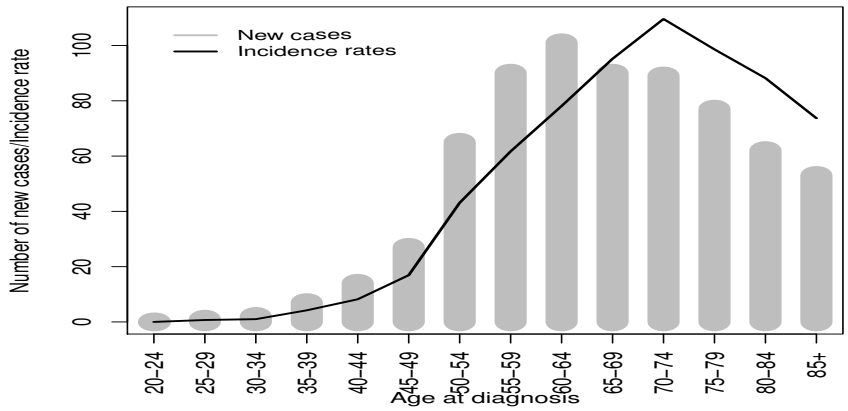


Figure 4.5: Average number of new breast cancer cases and age specific incidence rates in Norway from 2004 to 2008.

The risk of developing uterine cancer has increased steadily the last 50 years, from an average of 159 new cases in 1954-1958 to an average of 677 new cases in 2004-2008. The incidence rate has doubled in the same period, see Figure 4.6.

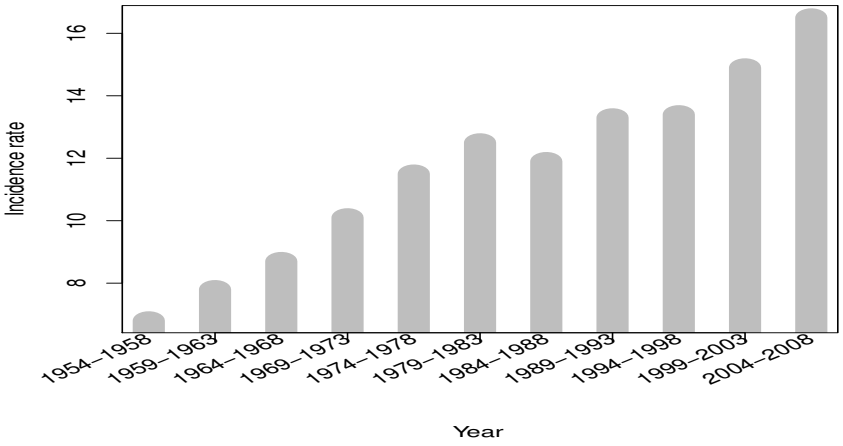


Figure 4.6: Average incidence rate per 100 000 person years, uterine cancer, 1954 to 2008.

Uterine cancer is the most common gynecologic cancer in Norway, with 716 out of 1565 new cases (approximately 46 %) in 2008. In many patients the disease is detected at an early stage, this provides good opportunities of healing. By the end of 2008, 8414 Norwegians had a uterine cancer diagnosis, approximately 45 % had

lived with the disease for more than 10 years. The probability of surviving uterine cancer in the period 2004 to 2008 was 83.2 %.

4.2.2 Current knowledge

With current knowledge one cannot directly point out the cause of uterine cancer. There exists knowledge about factors that may increase or decrease the risk of uterine cancer.

Kvåle, Heuch and Ursin [22] among other [26, 8] have studied the relationship between uterine cancer and reproductive factors. The risk of uterine cancer decreases with increasing parity and with increasing age at first and last birth. They also found a significant association with age at menarche and menopause, with the highest risk for women with long reproductive period, i.e. late menopause and/or early menarche.

A follow up of 1 million Norwegian women [6] showed that overweight/obesity is associated with increasing risk of uterine cancer, with a relative risk of 2.51 (95 % CI: 2.38 - 2.66) for obese women compared to lean women.

The Norwegian Cancer Society [5] mention hypertension, family history, prolonged exposure of estrogen and diabetes as other risk factors, they are not studied further in this analysis.

4.2.3 Uterine cancer in HUNT0

Approximately 1.5 % (934 women) of the participants was diagnosed with uterine cancer during follow-up. The average age at diagnosis was 67.7 years. The high average age can be explained by the fact that HUNT0 is a closed population, see Section 4.1.3.

Figure 4.7 shows a histogram of age by diagnosis. The plot seems to have a similar shape as Figure 4.5 which shows the age distribution for uterine cancer occurrences in Norway in 2004-2008. The incidence rate for each age can be seen in Figure 4.8. The figure shows a similar, but shifted, pattern compared to the incidence rate curve in Figure 4.5. The plot is smoothed by taking the average of the two closest ages on both sides.

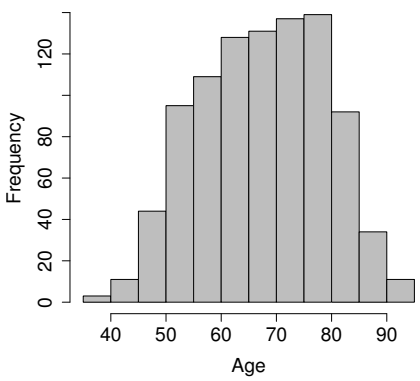


Figure 4.7: Histogram of age by uterine cancer diagnosis in HUNT0.

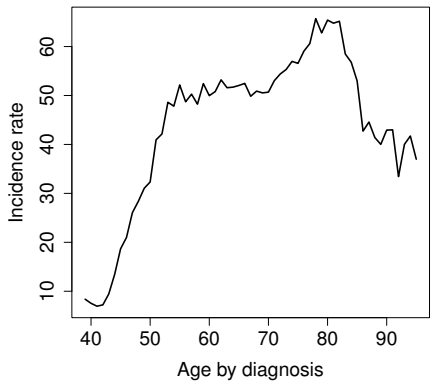


Figure 4.8: Incidence rate of uterine cancer per 100 000 person year.

4.3 Ovarian cancer

4.3.1 General information about ovarian cancer

Cancer Research UK [4], states that there were more than 225,000 new cases of ovarian cancer diagnosed worldwide in 2008. This is approximately 4% of all cancers diagnosed in women. In Norway, there were 457 new cases of ovarian cancer in 2008. This makes it the sixth most common cancer in Norwegian women, with approximately 3.8 % of all cancer cases. The cumulative risk of developing cancer by age 75 in Norway in 2008 was 1.3 %.

Figure 4.9 shows the average number of new ovarian cancer cases and average age specific incidence rates per 100 000 person years in Norway from 2004 to 2008 divided into age groups of five years. The figure shows that ovarian cancer is mainly a disease of older, postmenopausal women. The highest incidence rates are found among women aged 75-84. Over 85 % of all cases are diagnosed in women over 50 years.

In the recent years, the number of new ovarian cancer cases in Norway has stabilized with approximately 650-700 new cases each year. The incidence rate of ovarian cancer had a peak around year 1984-1993 and has since then decreased, (see Figure 4.10).

The symptoms of ovarian cancer are vague and the disease is therefore difficult to detect. For many patients the cancer has already spread when the cancer is diagnosed. This makes the treatment more difficult and the prognosis worse. The survival percent has been stable around 40 % since 1969. By the end of 2008, 4095

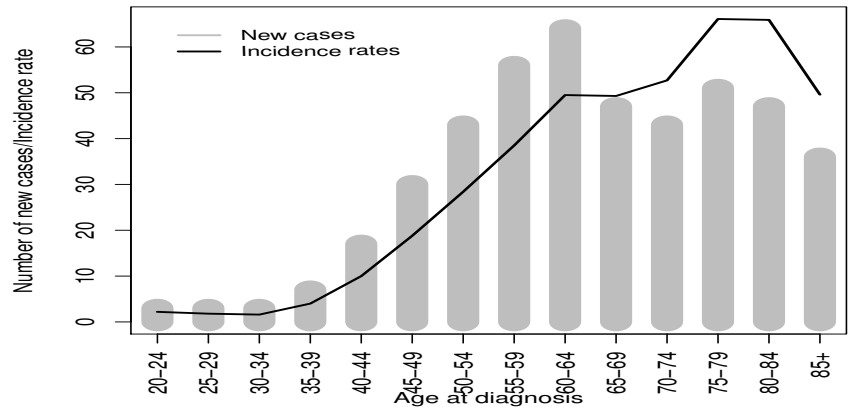


Figure 4.9: Average number of new ovarian cancer cases and age specific incidence rate per 100 000 person years in Norway from 2004 to 2008.

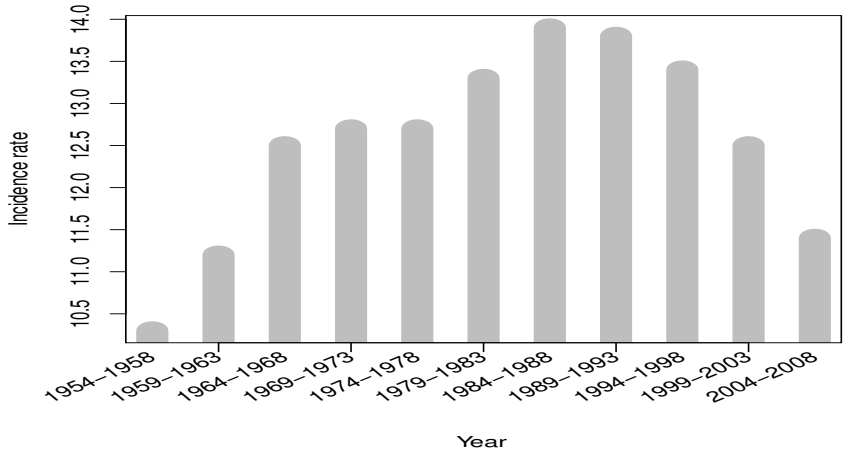


Figure 4.10: Average incidence rate per 100 000 person years, ovarian cancer, 1954 to 2008

women in Norway were alive with an ovarian cancer diagnosis, 2028 had lived with the cancer for more than 10 years.

4.3.2 Current knowledge

With current knowledge, it is incomprehensible why the normal cells in an ovary develop into cancer. There are, however, some known risk factors to be aware of. The risk of getting ovarian cancer increase with age as ovarian cancer often develops after menopause.

Reproductive factors and the risk of ovarian cancer has been a major area of research and most researchers agree that high parity is associated with a decreased risk of ovarian cancer, [13, 33, 21]. Previous reports of age at birth and ovarian cancer risk have been inconsistent. Kvåle, Heuch and Beral [21] states that age at first or last birth is not associated with ovarian cancer risk. Titus-Ernstoff [33] states the opposite, that age at first and last birth is associated with (reduced) ovarian cancer risk. The scientists are however consistent when it comes to age at menarche or menopause, it is not associated significantly with ovarian cancer risk.

The risk of getting ovarian cancer is found to be lower for women who have breast fed relative to those who have not, but the average duration of breast feeding is not found to be associated with ovarian cancer risk, [33].

Other known risk factors are tobacco use, oral contraceptives use, a previous cancer diagnosis, inherited gene mutation and hormone replacement therapy for menopause. These factors are not studied further in this analysis.

4.3.3 Ovarian cancer in HUNT0

During almost 50 years of follow up, approximately 1.4 % (843 women) of the cohort were diagnosed with ovarian cancer. The average age at diagnosis is approximately 68 years. Figure 4.11 and 4.12 shows histogram of number of ovarian cancer cases and incidence rate for each age, respectively. The figures show the same trend as the rest of Norway in 2004-2008, see Figure 4.9. The incidence rate plot is smoothed by taking the average of the relative age and the two nearest ages on both sides.

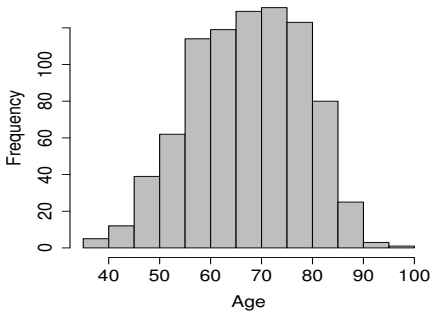


Figure 4.11: Histogram of age by ovarian cancer diagnosis.

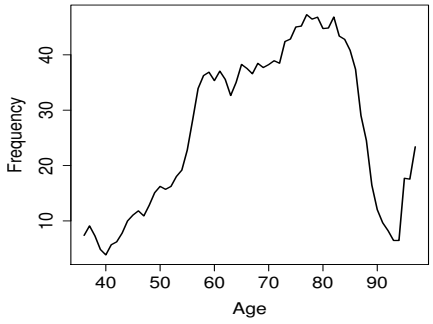


Figure 4.12: Incidence rate of ovarian cancer per 100 000 personyear.

Chapter 5

Explanatory data analysis

This chapter provides an insight into how each risk factor seems to influence the selected forms of cancer. A comparison of four different regression methods will also be performed. A more comprehensive analysis follows in Chapter 6 and 7.

5.1 Parity

As mentioned in Section 4.1.2, low parity is associated with increasing risk of breast, uterine and ovarian cancer. Figure 5.1 shows the average parity for each age, divided into four graphs; the risk set (Cox) and the cancer occurrences of breast, uterine and ovarian cancer. The graphs of the cancer cases are smoothed by taking the average of the relevant age and the four nearest ages. This applies to all similar plots in the rest of this chapter. In the plot it can be seen that the average parity for women diagnosed with cancer is evidently lower than the average of the risk set. The graphs of the cancer cases are not adjusted for the number of persons at risk at each age, hence the figure may give distorted impressions.

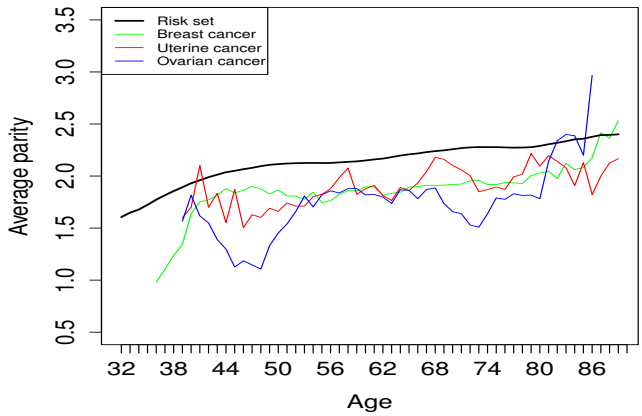
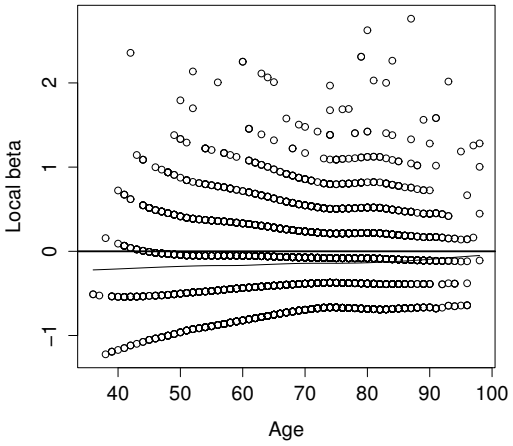
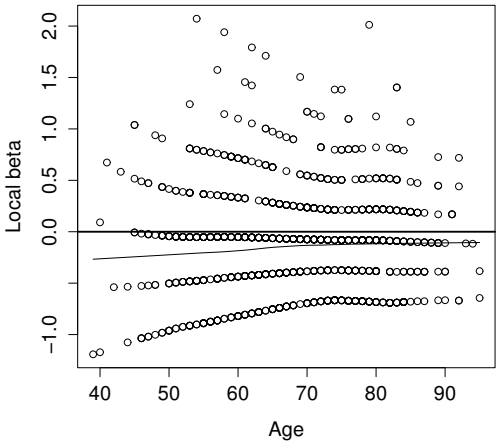


Figure 5.1: Average parity for the population and the cancer occurrences.

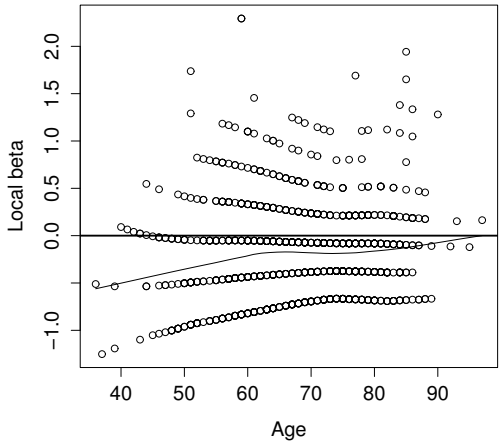
Local estimation of the coefficients in Approximate Cox regression for each competing event can be seen in Figure 5.2, with accompanying smoothing. The estimated coefficients are calculated by equation (2.28). The thick, black unbroken line indicates the zero-line. The plot shows that the coefficients, β_d, corresponding to breast and uterine cancer are approximately constant and equal to -0.1. Ovarian cancer seems to have a increasing β_d coefficient, a time dependent covariate or a division of the data set is therefore desirable, see Figure 5.2(c). However, there are apparently few ovarian cancer cases in the younger part of the cohort and the smooth parameter weights each point equivalently, the trend may therefore be ignored.



(a) Breast cancer



(b) Uterine cancer



(c) Ovarian cancer

Figure 5.2: Local estimation of coefficients in Approximate Cox regression, parity.

The risk set in Fine and Grey's proportional subdistribution hazard, (2.15), is extended with failures of competing events prior to the relevant time. Figure 5.3 shows the average parity for four risk sets; the risk set in Cox regression and the risk set for each of the three competing events in the proportional subdistribution hazard. Low parity increases the risk of breast, uterine and ovarian cancer. Hence, the average value of parity is reduced when the previous cancer occurrences of competing events are included in the risk set, as the figure shows. There are 3407 cases of breast cancer, 934 of uterine cancer and 843 of ovarian cancer. Clearly, the risk set for uterine and ovarian cancer will be more extended then the risk set of breast cancer, see figure 5.3. This may lead to a larger difference between the estimated coefficients in Cox proportional hazard and the proportional subdistribution hazard for uterine and ovarian cancer compared to breast cancer.

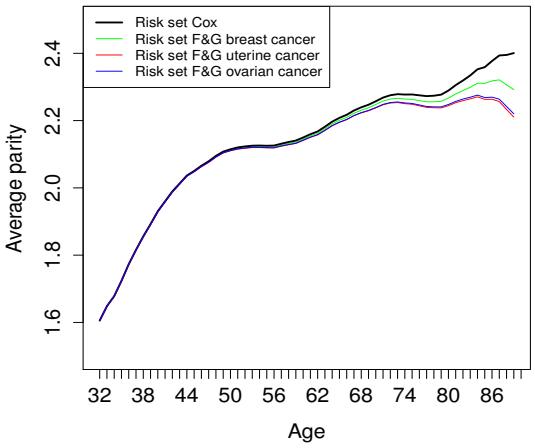


Figure 5.3: Average parity for four different risk sets.

Table 5.1 summarize the estimated coefficients in Cox proportional hazard model and the proportional subdistribution hazard model, with a corresponding standard deviation and p-value for parity. The estimates are not adjusted for any other covariates except age. The parameters are estimated by four methods; Cox regression, regression on the subdistribution hazard functions (see Section 2.4.2), Approximate Cox regression for general distribution of the covariates (see Section 2.6.5) and Approximate Fine and Grey regression. Each method seems to give approximately the same result, hence they are tantamount.

Overall, parity seems to affect the risk of breast, uterine and ovarian cancer. It is therefore recommended to include parity as a covariate in the model that describes the effect of covariates on the selected cancer types.

5.2 BMI

Figure 5.4 shows the average BMI for the cancer occurrences and the risk set (Cox) for each age. The figure indicates that the average BMI for women diagnosed with uterine cancer is considerably higher than the average BMI for the population at

	Parity				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	-0.106	0.899	0.011	<2e-16
	App. Cox	-0.097	0.908	0.010	< 2e-16
	Fine & Grey	-0.102	0.903	0.011	< 2e-16
	App. F & G	-0.097	0.908	0.010	< 2e-16
Uterine cancer	Cox	-0.104	0.901	0.021	7.43e-07
	App. Cox	-0.095	0.909	0.020	6.46e-07
	Fine & Grey	-0.097	0.908	0.021	4.15e-06
	App. F & G	-0.095	0.910	0.019	8.30e-07
Ovarian cancer	Cox	-0.171	0.843	0.023	2.72e-13
	App. Cox	-0.148	0.862	0.020	2.65e-13
	Fine & Grey	-0.165	0.848	0.023	1.98e-12
	App. F & G	-0.148	0.863	0.020	3.50e-13

Table 5.1: Estimated coefficients of parity with corresponding standard deviation and p-value computed by four methods.

risk. Low BMI in premenopausal women seems to be associated with increasing risk of breast cancer. The opposite result applies for postmenopausal women, where high BMI seems to increase the risk of breast cancer. The figure indicates that ovarian cancer risk is unrelated to BMI. However, the graphs of the cancer occurrences are not weighted for the number of women at risk at each age, the figure may therefore give misleading indications.

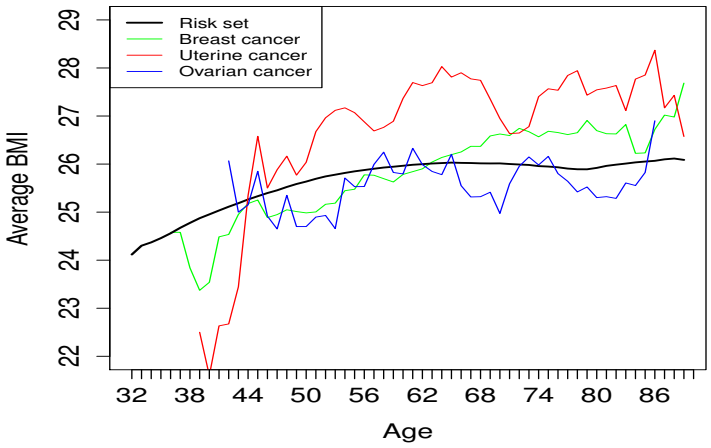
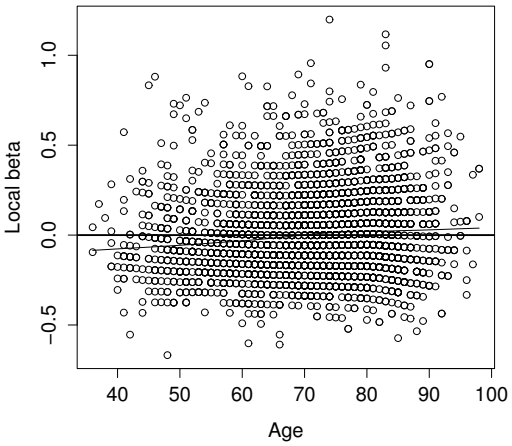
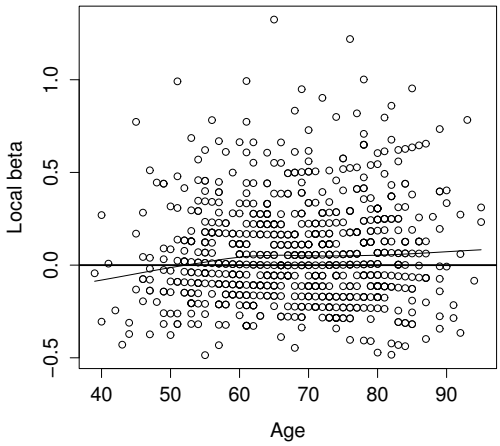


Figure 5.4: Average BMI for the population and cancer occurrences.

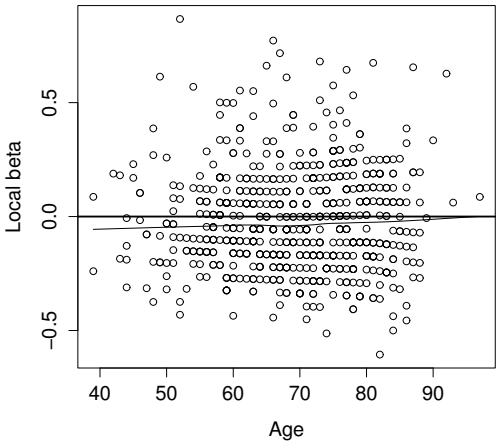
Local estimation of the regression parameters, β_d, in Approximate Cox regression (2.28) suggests that a time dependent covariate or separate analysis for pre and postmenopausal women is desirable when breast cancer is the cause of interest, see Figure 5.5(a). For uterine and ovarian cancer, a time independent covariate is sufficient as the smoothing lines seems constant, see Figure 5.5(b) and 5.5(c).



(a) Breast cancer



(b) Uterine cancer



(c) Ovarian cancer

Figure 5.5: Local estimation of coefficients in Approximate Cox regression, BMI.

Figure 5.6 shows the average BMI of the population at risk and the risk set in Fine and Grey's method, (2.15). The figure shows that the four risk sets are almost equal. This can be explained by the fact that BMI has the opposite effect on ovarian and breast cancer risk compared to uterine cancer risk.

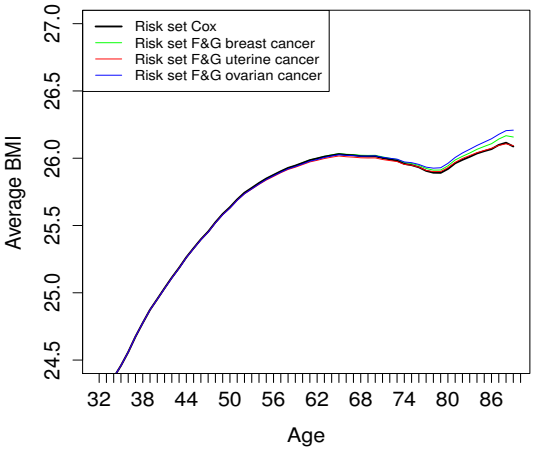


Figure 5.6: Average BMI for four different risk sets.

Table 5.2 shows the estimated coefficients of BMI with standard deviation and p-value for each competing event, computed by the four methods described earlier. The estimates are not adjusted for any of the other explanatory variables except age. Regression on the cause specific hazard functions and the subdistribution hazard functions gives approximately the same result.

Overall, the risk of cancer seems to be related to BMI, it is therefore recommended to include BMI as a covariate in the optimal model. Separate analysis for pre and postmenopausal women is also recommended.

	BMI				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	0.019	1.019	0.004	2.62e-05
	App. Cox	0.019	1.020	0.005	2.53e-05
	Fine & Grey	0.018	1.019	0.004	3.64e-05
	App. F & G	0.019	1.019	0.005	2.92e-05
Uterine cancer	Cox	0.064	1.066	0.008	<2e-16
	App. Cox	0.073	1.076	0.009	<2e-16
	Fine & Grey	0.065	1.067	0.008	<2e-16
	App. F & G	0.073	1.076	0.009	<2e-16
Ovarian cancer	Cox	-0.016	0.985	0.010	0.112
	App. Cox	-0.015	0.985	0.010	0.113
	Fine & Grey	-0.015	0.985	0.010	0.116
	App. F & G	-0.015	0.985	0.010	0.109

Table 5.2: Estimated coefficients of BMI with corresponding standard deviation and p-value computed by four methods.

5.3 Age at first birth

The average age at first birth for the population at risk and the cancer occurrences can be seen in Figure 5.7. Women with a breast cancer diagnosis seem to have a higher average age at first birth than the population. The opposite effect applies to uterine cancer, where the average age at first birth is lower than the population. Ovarian cancer risk seems to be unrelated to age at first birth.

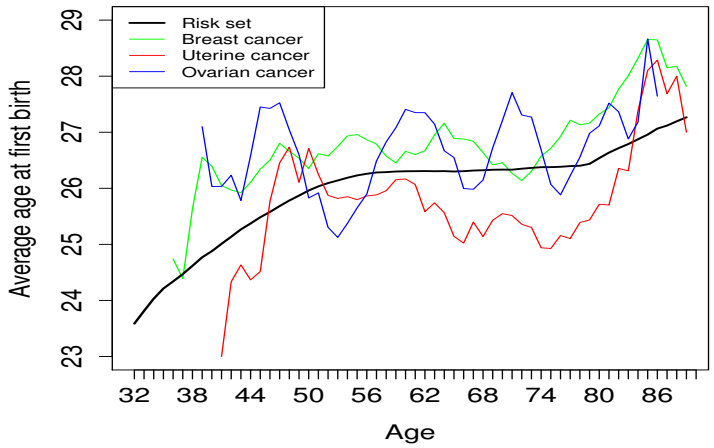


Figure 5.7: Average age at first birth for the population and the cancer occurrences.

Plots of the local β_d parameters in Approximate Cox regression indicate that age at first birth is a time-independent covariate for all the competing events, the plots are not included here.

Table 5.4 shows the estimated coefficients in Cox proportional hazard and proportional subdistribution hazard with standard deviation and p-value for age at first birth, computed by the four methods described earlier. The estimates are not adjusted for any explanatory variables except age. The table suggests that the four estimation methods gives approximately the same result and also coincide with the trends mentioned above.

5.4 Age at last birth

By comparing the average age at last birth for the the population at risk with the average age at last birth for women diagnosed with cancer, (see Figure 5.8), it seems likely that early age at last birth increases the risk of uterine and ovarian cancer, but is unrelated to breast cancer risk.

	Age at first birth				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	0.024	1.024	0.004	1.44e-09
	App. Cox	0.024	1.025	0.004	4.10e-10
	Fine & Grey	0.025	1.025	0.004	7.71e-10
	App. F & G	0.024	1.025	0.004	4.81e-09
Uterine cancer	Cox	-0.028	0.972	0.008	0.0006
	App. Cox	-0.028	0.973	0.008	0.0005
	Fine & Grey	-0.029	0.972	0.008	0.0005
	App. F & G	-0.028	0.973	0.008	0.0004
Ovarian cancer	Cox	0.016	1.016	0.008	0.06
	App. Cox	0.016	1.016	0.008	0.05
	Fine & Grey	0.016	1.016	0.008	0.05
	App. F & G	0.016	1.016	0.008	0.07

Table 5.3: Estimated coefficients of age at first birth with corresponding standard deviation and p-value computed by four methods.

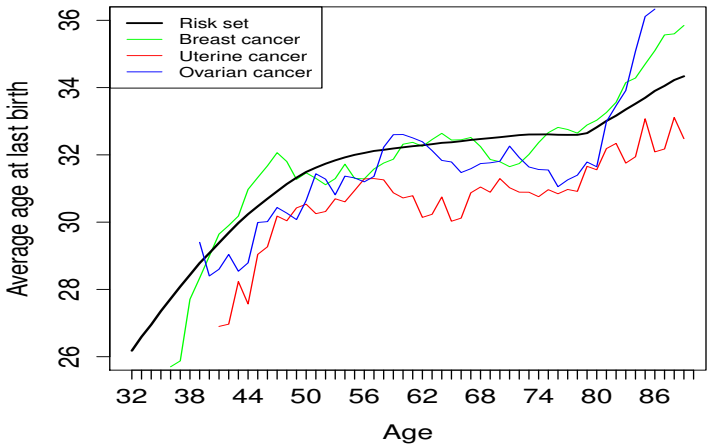


Figure 5.8: Average age at last birth for the population and the cancer occurrences.

Local estimation of the β_d coefficients in Approximate Cox regression show constant negative parameters when uterine and ovarian cancer are the cause of interest. Age at last birth seems to be unrelated to breast cancer risk, with an estimated coefficient approximately equal to 0. The plots are not included here.

Table 5.4 support the conclusions drawn from Figure 5.8, that early age at last birth increases the risk of uterine and ovarian cancer, but is unrelated to breast

cancer risk. Each of the four estimation methods seems to give approximately the same result.

	Age at last birth				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	0.001	1.001	0.004	0.863
	App. Cox	-0.0005	0.999	0.004	0.843
	Fine & Grey	0.002	1.002	0.004	0.602
	App. F & G	0	1	0.004	0.939
Uterine cancer	Cox	-0.048	0.953	0.007	1.81e-11
	App. Cox	-0.048	0.953	0.007	7.03e-12
	Fine & Grey	-0.046	0.955	0.007	9.97e-11
	App. F & G	-0.048	0.953	0.007	8.06e-12
Ovarian cancer	Cox	-0.018	0.982	0.008	0.019
	App. Cox	-0.018	0.982	0.008	0.014
	Fine & Grey	-0.016	0.984	0.008	0.038
	App. F & G	-0.018	0.982	0.008	0.015

Table 5.4: Estimated coefficients of age at last birth with corresponding standard deviation and p-value computed by four methods.

5.5 Age at menarche

The average age at menarche for women diagnosed with cancer seems to be lower than the average age at menarche for the population, see Figure 5.9.

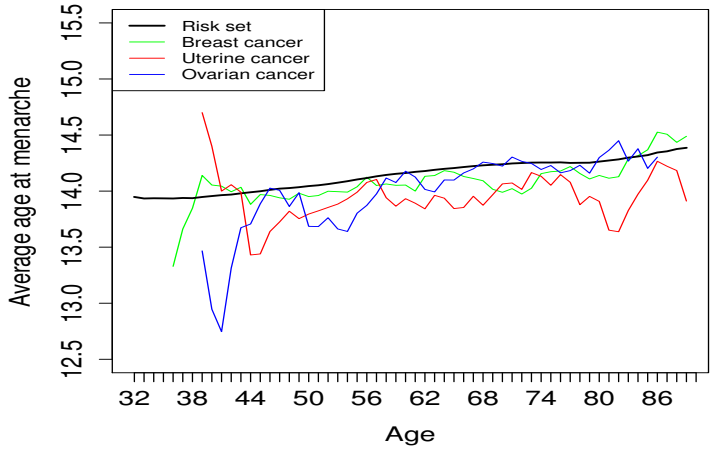


Figure 5.9: Average age at menarche for the population and the cancer occurrences.

Plots of the local β_d parameters in Approximate Cox regression appear to be constant for all competing events, hence a time dependent covariate is not necessary. The plots are not included here.

Table 5.5 shows the estimated coefficients with standard deviation and p-value computed by the four methods mentioned earlier. The result given in the table coincide with Figure 5.9, early age at menarche decreases the risk of cancer. The estimates are not adjusted for any explanatory variables, except age.

	Age at menarche				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	-0.051	0.951	0.013	5.46e-05
	App. Cox	-0.052	0.950	0.012	3.42e-05
	Fine & Grey	-0.047	0.954	0.013	0.0002
	App. F & G	-0.051	0.950	0.012	3.99e-05
Uterine cancer	Cox	-0.126	0.882	0.024	2.34e-07
	App. Cox	-0.123	0.884	0.024	1.89e-07
	Fine & Grey	-0.121	0.887	0.024	6.85e-07
	App. F & G	-0.123	0.885	0.024	2.16e-07
Ovarian cancer	Cox	-0.042	0.959	0.025	0.100
	App. Cox	-0.042	0.959	0.025	0.150
	Fine & Grey	-0.036	0.965	0.025	0.155
	App. F & G	-0.041	0.960	0.025	0.099

Table 5.5: Estimated coefficients of age at menarche with corresponding standard deviation and p-value computed by four methods.

5.6 Age at menopause

Figure 5.10 indicates that late age at menopause increases the risk of breast and uterine cancer, but not ovarian cancer.

Plots of local β_d parameters in Approximate Cox regression (2.28) show a constant trend for age at menopause, this applies to all the competing events. Hence, a time dependent covariate is not necessary. The plots are not included here.

Table 5.6 shows the estimated coefficients in proportional cause specific hazard and proportional subdistribution hazard with standard deviation and p-value calculated by the four methods mentioned earlier. The table supports the findings from Figure 5.10, that late age at menopause increases the risk of breast and uterine cancer.

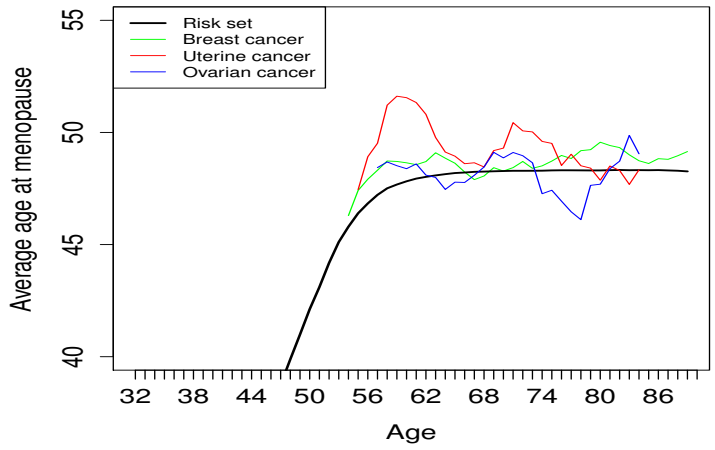


Figure 5.10: Average age at menopause for the population and the cancer occurrences.

	Age at menopause				
	Method	Coef	Exp(coef)	S.D.(coef)	P-value
Breast cancer	Cox	0.032	1.033	0.009	0.0002
	App. Cox	0.030	1.030	0.008	0.0002
	Fine & Grey	0.032	1.033	0.009	0.0002
	App. F & G	0.031	1.031	0.008	0.0001
Uterine cancer	Cox	0.059	1.061	0.022	0.007
	App. Cox	0.052	1.053	0.019	0.006
	Fine & Grey	0.060	1.061	0.022	0.007
	App. F & G	0.054	1.055	0.019	0.006
Ovarian cancer	Cox	-0.004	0.996	0.016	0.802
	App. Cox	-0.004	0.996	0.016	0.798
	Fine & Grey	-0.004	0.996	0.016	0.790
	App. F & G	-0.002	0.998	0.016	0.886

Table 5.6: Estimated coefficients of age at menopause with corresponding standard deviation and p-value computed by four different methods.

5.7 Other

Variables like lactation, abortion, marital status, occupation and residence have also been investigated. These variables show no clear trend and are therefore not commented any further in this chapter.

Chapter 6

Regression on the cause specific hazard functions

The aim this chapter is to find a common model that best describes the relationship between the various explanatory variables and the risk of breast, ovarian or uterine cancer in the presence of the competing events. A thorough analysis of each cancer type is given in Chapter 7.

6.1 Model description

The dataset was originally designed to investigate the circumstances surrounding breast cancer in women. It will be well suited for analysis of cancer in the genital organs because of similar risk factors. The material is unique due to many explanatory variables and observations.

The explanatory data analysis revealed that Cox regression, Approximate Cox regression, "Fine and Grey regression" (see Section 2.4.2) and Approximate Fine and Grey regression gives approximately the same result. Cox proportional hazards model will therefore be used to analyse the relationship between the explanatory variables and the risk of breast, ovarian or uterine cancer risk in the presence of the competing events.

It is important to notice that this is not always the case. If there were more breast, uterine or ovarian cancer cases, the risk set in Fine and Grey's method would be larger, and a (possible) increased difference between the estimated coefficients in Cox proportional hazard, (2.13), and the proportional subdistribution hazard, (2.15), would follow.

A large data set is not only an advantage, many of the explanatory variables have a large number of missing values. Women with missing values are not taken out of the analysis, but ignored by the statistical software R [2] in analyses where the

values are missing. In HUNT0, information regarding reproductive factors will be incomplete for premenopausal women, and a large number of missing values arise. For instance, age at first or last birth and parity lack information on approximately 1/5th of the participants. A comparison of different models will therefore be difficult since the data sets differ.

From the explanatory data analysis and current knowledge in Chapter 4, it is known that reproductive factors, BMI and reproductive period are associated with increasing or decreasing risk of breast, ovarian or uterine cancer. These factors will be emphasized in a final model in spite of many missing values.

A brief analysis shows that parity + age at first birth and age at last birth are highly correlated, and that age at first birth is non-significant as an explanatory variable when age at last birth and parity are included in the model. This applies to all competing events. Age at first and last birth have the same missing values. This means that including just one of them will not change the dataset. Age at first birth will therefore not be included in the selected model to describe the diseases in spite of the clear trend observed in the explanatory data analysis. Due to the large amount of missing values (4/5th of the participants), age at menopause will not be included in the final model but rather described in detail in the next chapter.

From the explanatory data analysis it is known that demographic data, occupation and marital status are non-significant as explanatory variables for any of the competing events. These covariates will therefore not be included in the investigated model.

The selected variables in the optimal model are parity, age at last full term pregnancy, BMI and age at menarche. All of them are treated as continuous covariates. Some of the variables will probably be unnecessary or non-significant in describing the risk of breast, ovarian or uterine cancer. However, all of them are included because they are important for at least one of the three competing events. Interaction terms between the covariates have been tested out, these are found to be highly non-significant and are therefore not included in the model.

The cause specific hazard functions from equation (2.13) are the fitted models. As mentioned in Section 3.2.1, age is chosen as the relevant time parameter. Alternatively, time after 1961 could have been used with age included as an explanatory variable. Since the women enter the survey at different ages, left truncation is necessary. The "follow up" time, t, is calculated as the time from 01.01.1961 until date of death, cancer diagnosis, emigration or termination date, whichever comes first.

Some women received a cancer diagnosis or died in 1961. Statistical analysis in R [2] requires a positive difference between entry date and termination date. Hence,

all women that died or got a breast, uterine or ovarian cancer diagnosis in 1961 are registered with their age in 1962 as "out-age". This applies to 27 out of 5184 women, and the error may therefore be neglected. Alternatively, a continuous time parameter could have been used. A discrete time parameter is more appropriate and therefore desirable.

With covariates, the model can be written

$$\lambda_d(t;\mathbf{X}) = \lambda_{d0}(t)\exp\Big(\beta_{d1}parity + \beta_{d2}agelast + \beta_{d3}bmi + \beta_{d4}menarche\Big); \qquad d = 1,2,3. \tag{6.1}$$

Table 6.1 shows the estimated coefficients, relative risk, 95 % confidence interval for relative risk and P-value from Cox regression on model (6.1), with breast, uterine or ovarian cancer as endpoint. The function *coxph* from the library *survival* in the statistical software R [2] is used to obtain the results.

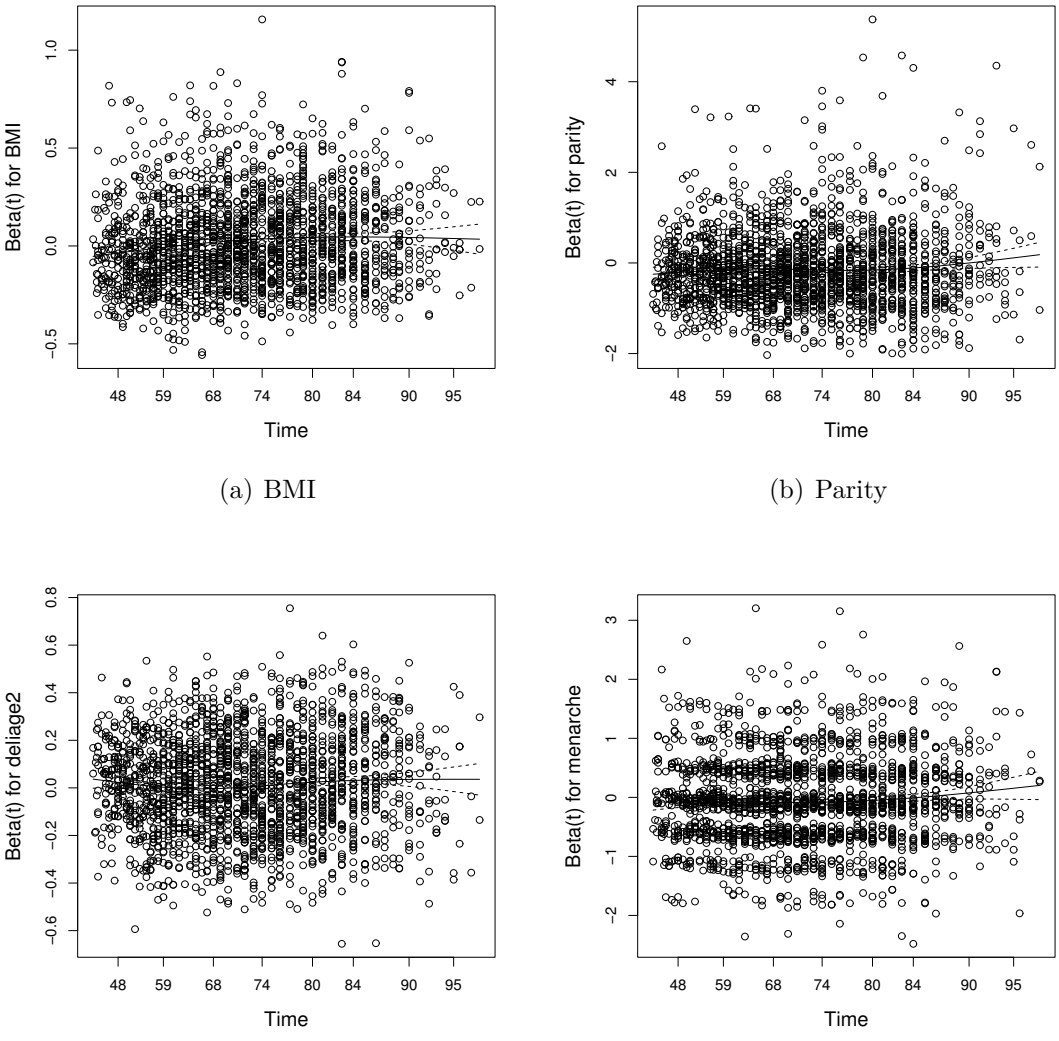
	Parameter estimates					
	Explanatory variable		Coef	RR	95 % CI for RR	P-value
Breast cancer	Parity	β_{11}	-0.141	0.869	(0.838, 0.901)	5.7e-14
	Age at last birth	β_{12}	0.013	1.013	(1.004, 1.022)	0.004
	BMI	β_{13}	0.020	1.020	(1.010, 1.031)	0.0002
	Age at menarche	β_{14}	-0.059	0.942	(0.913, 0.974)	0.0003
Uterine cancer	Parity	β_{21}	-0.055	0.946	(0.882, 1.015)	0.12
	Age at last birth	β_{22}	-0.050	0.951	(0.935, 0.968)	5.71e-08
	BMI	β_{23}	0.068	1.071	(1.052, 1.090)	3.49e-14
	Age at menarche	β_{24}	-0.094	0.910	(0.856, 0.969)	0.003
Ovarian cancer	Parity	β_{31}	-0.152	0.859	(0.791, 0.932)	0.0003
	Age at last birth	β_{32}	-0.008	0.992	(0.973, 1.011)	0.41
	BMI	β_{33}	0	1	(0.977, 1.023)	0.99
	Age at menarche	β_{34}	-0.035	0.966	(0.902, 1.034)	0.32

Table 6.1: Estimated coefficients, relative risk, 95% confidence interval for relative risk and p-value for the covariates in model (6.1).

The relative risk, RR, for a continuous covariate is the effect of one unit increase while all other parameters are kept constant. A relative risk approximately equal to one implies a non-significant term.

In Cox proportional hazards model it is a necessity that the cause specific hazard functions are proportional. A Schoenfeld residual plot can be used to check for lack of fit over time for continuous covariates. The residuals should not indicate any

pattern and should have a local mean around 0. Figure 6.1 shows Schoenfeld residuals when breast cancer is the cause of interest. Similar plots for uterine and ovarian cancer shows no evident trend, and are therefore omitted. The model assumptions are fulfilled and model (6.1) is the final model.



(a) BMI

(b) Parity

(c) Age at last birth

(d) Age at menarche

Figure 6.1: Schoenfeld residuals for the four covariates in model (6.1), breast cancer.

Chapter 7

Medical results

7.1 Breast cancer

3407 out of 61457 participants were diagnosed with breast cancer during follow-up.

Analysis of HUNT0 data shows that the risk of breast cancer decreases with increasing parity, with a relative risk of 0.90 (p-value: $<$ 2e-16). This means that each additional birth reduces the risk of breast cancer with 10 %. If parity is treated as a categorical covariate, with states 0-5 births (group 5 is 5 births or more), the relative risk for a woman with 5 or more births compared to a nulliparous woman is 0.52 (p-value: 2.11e-15). Figure 7.1 shows the cumulative incidence curves for breast cancer divided into the 6 "birth-groups". The function *etmCIF* in the package *etm* used by the statistical software R [2] is used to estimate the non-parametric cumulative incidence functions from competing risks data in the presence of left truncation and right censoring. Clearly, many births decreases the risk of breast cancer compared to few births. Previous studies show similar results, see Section 4.1.2.

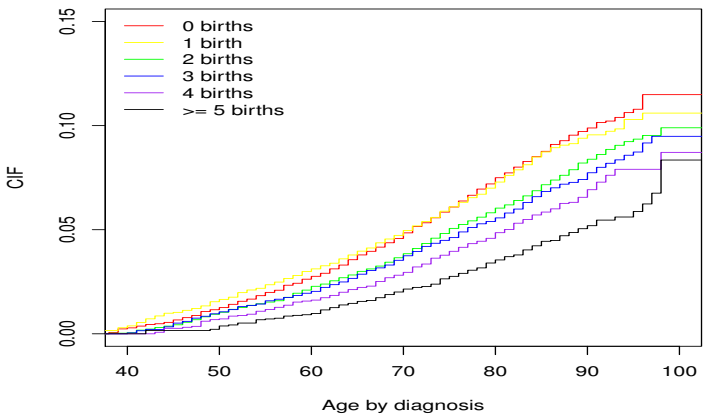


Figure 7.1: Cumulative incidence functions for breast cancer divided into six parity groups.

Age at first birth is found strongly significant as an explanatory variable with a relative risk of 1.025 (p-value: 1.44e-09). This means that each year increase in age at first birth increases the risk of breast cancer with 2.5 %. The relative risk of breast cancer for women with the first birth later than the age of 30 compared to first birth before the age of 20 is 1.45 (p-value: 8.91e-06). Obviously early age at first birth appear protective against breast cancer compared to late age at first birth. Figure 7.2 shows the cumulative incidence functions for breast cancer divided into four first birth categories. The figure emphasizes the result mentioned above. Previous studies show similar results, see Section 4.1.2.

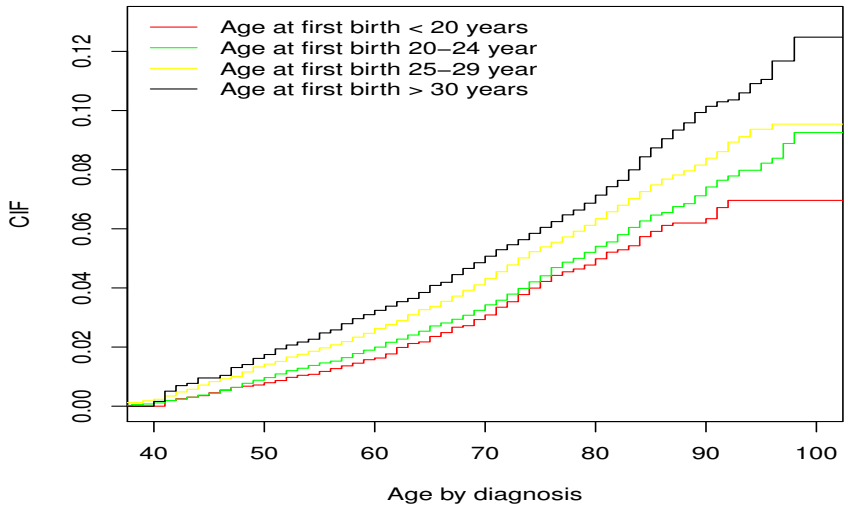


Figure 7.2: Cumulative incidence functions for breast cancer divided into four age at first birth groups.

Separate analysis for pre and postmenopausal women shows that age at first birth is non-significant as an explanatory variable for postmenopausal women. For premenopausal women, the risk of getting breast cancer increases with approximately 3 % for each additional year in age at first birth (p-value: 1.27e-08).

As mentioned in Section 4.1.2, the relationship between age at first birth, age at last birth and parity is very complex and the variables are strongly correlated. The significant association between age at first birth and breast cancer risk is weakened when adjusted for parity. Further adjustment for age at last birth makes age at first birth non-significant as an explanatory variable.

Age at last birth showed initially no association with breast cancer risk and is not significant as an explanatory variable. After additional adjustment for parity, a significant association emerges with a relative risk of 1.02 for each additional year

in age at last birth (p-value: 3.98e-05). The relative risk of breast cancer for women with the last birth after the age of 40 compared to younger than 20 years is 1.30 (p-value: 0.009). Early last birth is protective against breast cancer compared to late last birth. Further adjustment for BMI emphasizes the result.

A significant increase in breast cancer risk is observed with increasing age at menopause, with a relative risk of 1.03 (p-value: 0.0002). This corresponds to an increasing risk of breast cancer by 3 % for each additional year in age at menopause. Adjustments for parity and age at last birth does not influence this estimate considerably.

A significant association between breast cancer risk and age at menarche is observed with a relative risk of 0.95 (p-value: 0.0005). This estimate corresponds to an average decrease in breast cancer risk with 5 % for each year increase in age at menarche. The effect of age at menarche on breast cancer risk is still obtained in the analysis after additional adjustment for parity, age at last birth and BMI.

The risk of getting breast cancer increases with increasing BMI, with a relative risk of 1.02 (p-value: 2.62e-05) for each unit increase in BMI. Obese women have a relative risk of 1.25 (p-value: 1.44e-05) compared to lean women. Figure 7.3 shows the cumulative incidence functions for breast cancer. The graph is divided into lean, overweight and obese women. The figure shows that obese women have the lowest risk of breast cancer before menopause, and the highest risk after menopause.

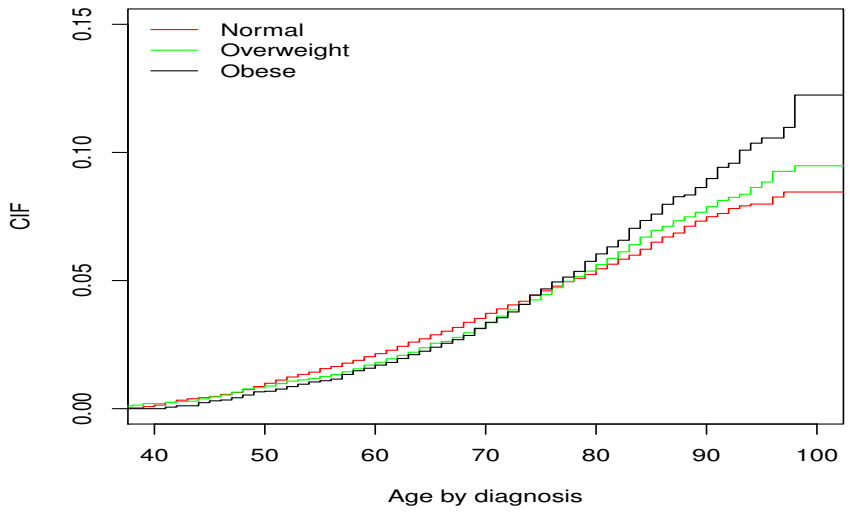


Figure 7.3: Cumulative incidence functions for breast cancer, divided into normal, overweight and obese women.

The explanatory data analysis suggests that high BMI have different effect on breast

cancer risk for pre and postmenopausal women. Separate analysis for women that have undergone menopause (or passed 55 years) is therefore desirable. The risk of getting breast cancer for postmenopausal women increases with increasing BMI (RR per unit increase in BMI = 1.03, p-value: 0.0005). Compared to postmenopausal women with normal BMI, overweight and obese women have a relative risk of 1.19 (p-value: 0.06) and 1.38 (p-value: 0.002), respectively. Adjustments for parity enhance the result. Further adjustment for age at menarche and age at last birth makes BMI non-significant as an explanatory variable and the trend vanish. In spite of the indications in the explanatory data analysis, BMI have the same effect on pre and postmenopausal women. In premenopausal women, the relative risk for each unit increase in BMI is 1.02 (p-value: 0.0005). The trend seems to be weaker for premenopausal women compared to postmenopausal women. The relative risk for obese premenopausal women compared to lean premenopausal women is 1.27 (p-value: 9.76e-5). No evident association is found for overweight women.

In Section 4.1.2 lactation was mentioned as a possible risk factor for breast cancer. In this study, lactation is not found to influence the risk of breast cancer.

Women that have undergone an abortion have a decreased risk of getting breast cancer compared to women that have not (RR = 0.91, p-value = 0.04). Adjustment for parity makes the association to vanish.

7.2 Uterine cancer

Among the 61457 women who attended the screening program, 934 cases of uterine cancer were diagnosed.

The risk of uterine cancer decreases with increasing parity, with a relative risk of 0.9 for each additional birth (p-value: 7.43e-07). The association is strengthened when adjusting for age at first birth. A woman with five or more births have a relative risk of 0.52 (p-value: 3.06e-05) compared to a nulliparous woman. Figure 7.4 shows the cumulative incidence functions for uterine cancer divided into six groups, 0-4 births and 5 or more births. The figure shows how multiparous women have a decreased risk of getting uterine cancer. Current knowledge on uterine cancer risk show similar trends, see Section 4.2.2.

From the explanatory data analysis and Section 4.2.2 it is known that the risk of uterine cancer decreases with increasing age at first birth. The relative risk for each year increase in age at first birth is 0.97 (p-value: 0.0006). Women with their first birth after the age of 30 years have approximately 41 % reduced risk of developing uterine cancer compared to women that gives the first birth before the age of 20 (p-value: 0.0005). Adjusting for parity strengthen the result.

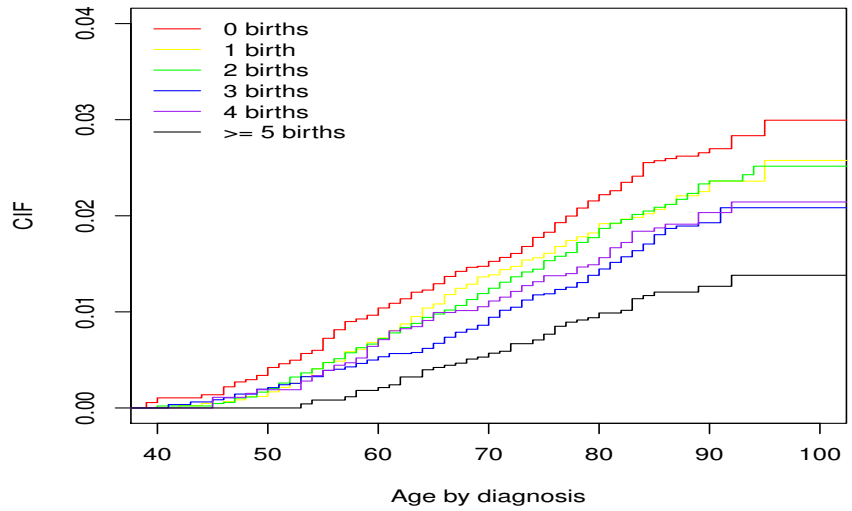


Figure 7.4: Cumulative incidence functions for uterine cancer divided into six parity groups.

Late age at last birth decreases the risk of uterine cancer with approximately 5 % for each additional year (p-value: 1.81e-11). The relative risk of uterine cancer for women with their last birth after the age of 40 compared to the last birth before the age of 25 is 0.39 (p-value: 1.20e-06).

As mentioned earlier, the relationship between age at first and last birth is very complex. Both age at first birth and parity are significant as an explanatory variable alone, however, adjusting for age at last birth makes both of them non-significant.

The risk of getting uterine cancer increases with approximately 7 % for each additional unit increase in BMI (p-value: $<$ 2e-16). Compared to lean women, overweight and obese women have a relative risk of 1.25 (p-value: 0.008) and 1.95 (p-value: 1.40e-12), respectively. The association is strengthened by adjusting for parity. Figure 7.5 show cumulative incidence functions when uterine cancer is the cause of interest. The figure shows that obese women have about twice the risk of getting uterine cancer compared to lean women. The same trend has been found in previous studies, see Section 4.2.2.

High age at menopause is found to increase the risk of uterine cancer, with a relative risk of 1.06 (p-value: 0.007) for each additional year. This means that each extra year increases the risk of uterine cancer with approximately 6 % compared to the year before. The association is weakened when adjusting for BMI or parity, and strengthened when adjusting for age at last birth. Previous studies have reached similar results, see Section 4.2.2. Figure 7.6 show the cumulative incidence functions

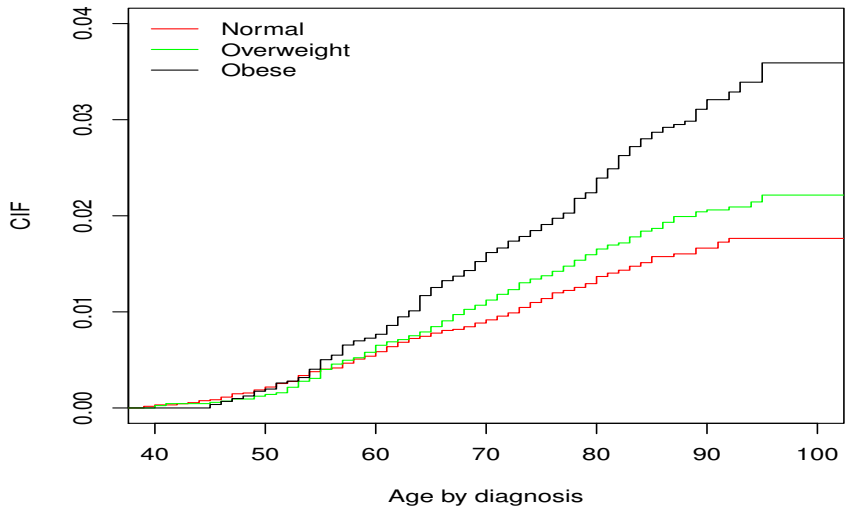


Figure 7.5: Cumulative incidence functions for uterine cancer divided into normal, overweight and obese women.

for uterine cancer divided into groups of age at menopause. The graph emphasizes the result mentioned above.

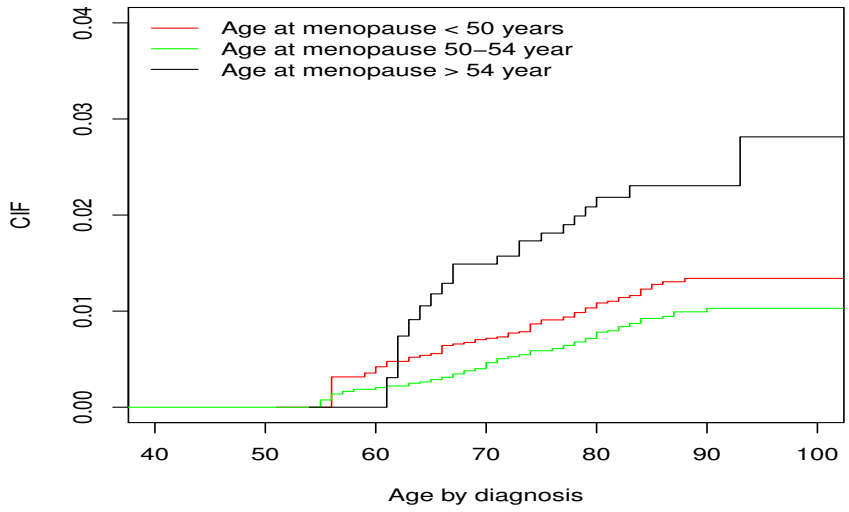


Figure 7.6: Cumulative incidence functions for uterine cancer divided into groups of age at menopause.

Each year increase in age at menarche is found to decrease the risk of uterine cancer with approximately 12 % (p-value: 2.34e-07). The result is unchanged when adjusting for parity and age at last birth. A short reproductive period seems to be protective against uterine cancer.

Analyses of marital status show a decreasing risk of uterine cancer for married women compared to non-married women. This trend vanish when adjusting for parity, as it is reasonable to assume that single women are childless. Lactating women have a decreased risk of uterine cancer compared to non-lactating women. This trend is not significant and vanish when adjusting for other covariates. Neither occupation, residence, county nor abortion show any specific trends.

7.3 Ovarian cancer

Among 61457 women, 843 cases of ovarian cancer were diagnosed during follow up.

Analysis of HUNT0 data shows that the risk of ovarian cancer decreases with increasing parity, with a relative risk of 0.84 for each additional birth (p-value: 2.72e-13). This means that each extra birth reduces the risk of getting ovarian cancer by 16 %. The risk of developing ovarian cancer for a woman with five or more births compared to a nulliparous woman is 0.41 (p-value: 1.54e-07). Adjusting for BMI, age at last birth and age at menarche does not affect the result noticeably. Figure 7.7 shows the cumulative incidence functions when ovarian cancer is the cause of interest. The graph is separated into six different groups, 0-4 births and 5 or more births. The graph emphasizes the result mentioned above. The result is similar to current knowledge on ovarian cancer risk, see Section 4.3.2. The explanatory data analysis suggests separate analyses for pre and postmenopausal women. Further analysis shows that this is not necessary, as the estimated coefficients seem constant for all ages.

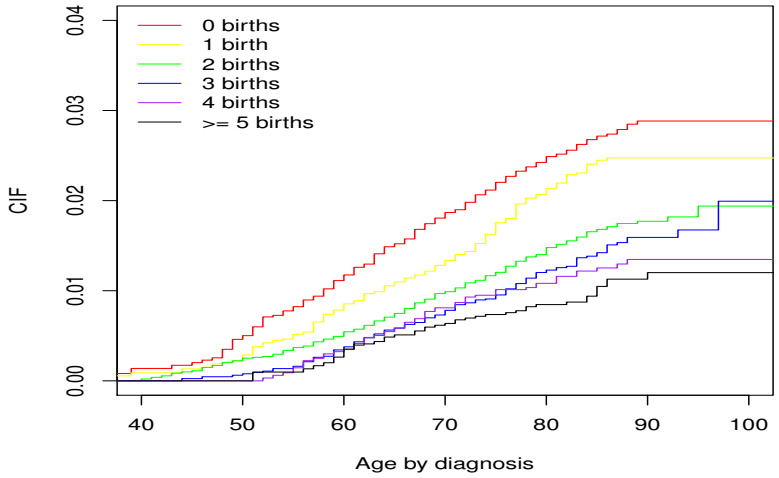


Figure 7.7: Cumulative incidence functions for ovarian cancer divided into six parity groups.

Initially, BMI was found to be a non-significant explanatory variable for uterine cancer risk, see Section 5.2. However, when BMI is treated as a categorical variable (see Section 3.2.4), a significant association between obesity and ovarian cancer risk emerge with a relative risk of 0.77 for obese women compared to lean women (p-value: 0.03). The same decreasing trend can be seen for overweight women, with a relative risk of 0.93, this trend is not significant. Figure 7.8 shows the cumulative incidence functions for ovarian cancer divided into normal, overweight and obese women. The figure shows a reduced risk of getting ovarian cancer for obese women compared to overweight and lean women.

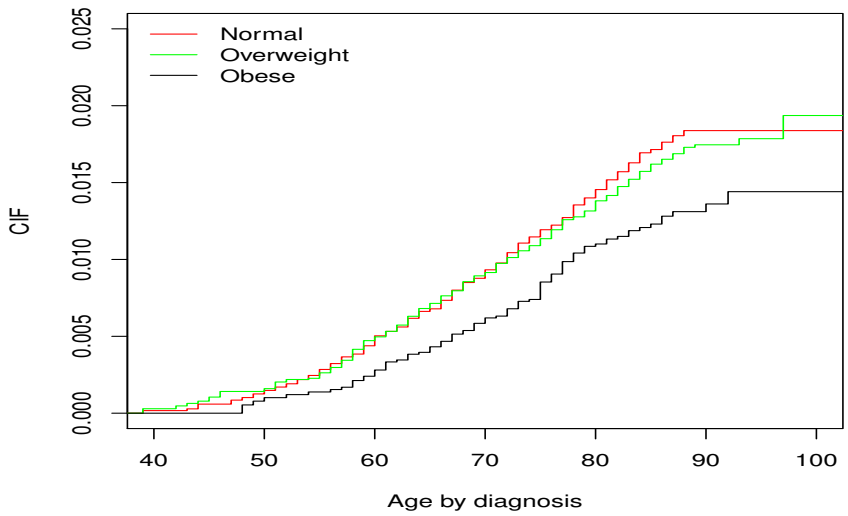


Figure 7.8: Cumulative incidence functions for ovarian cancer divided into normal, overweigh and obese women.

Further analysis of BMI shows that for postmenopausal women, high BMI decreases the risk of getting ovarian cancer with approximately 6 % for each unit increase in BMI (p-value: 0.003). Adjustment for parity does not alter this result. The relative risk of getting ovarian cancer for obese postmenopausal women compared to lean women is 0.42 (p-value: 0.002). The same trend can be seen for overweight women, however, the trend is not significant. BMI does not show the same trend and is not significant as an explanatory variable for premenopausal women (p-value: 0.85). Hence, overweight seems to be protective against ovarian cancer for postmenopausal women, but not premenopausal.

Each additional year in age at last birth decreases the risk of ovarian cancer with approximately 2 % (p-value: 0.02). The association vanish when adjusting for parity. Separate analyses for pre and postmenopausal women shows that postmenopausal women with late last birth have a decreased risk of getting ovarian cancer (RR:

0.97 for each unit increase, p-value: 0.02). No trend is observed for premenopausal women.

Age at first birth showed initially a nearly significant association to ovarian cancer risk with a relative risk of 1.02 for each year increase in age at first birth (p-value: 0.06). The trend vanish when adjusting for parity.

Early age at menarche is associated with a decreasing risk of ovarian cancer (RR: 0.96 for each unit increase, p-value: 0.1), however, the trend is not significant when adjusting for parity. Age at menopause and demographic variables have no association to ovarian cancer risk in this study. The same conclusion is drawn in other studies, see Section 4.3.2.

Overall, adjusting for parity makes all the other explanatory variables unnecessary in describing ovarian cancer risk. The only exception is BMI and postmenopausal women.

Chapter 8

Concluding remarks

The theory of competing risks is needed to obtain realistic results when a unit can fail due to several failure causes. In this thesis, it has been shown how the theory of competing risks can be used in a medical study with breast, uterine and ovarian cancer as the competing events.

Both regression on the cause specific hazard functions and the subdistribution hazard functions have been used to identify possible risk factors for the competing events. Regression on the cause specific hazard has been implemented by Cox regression and Approximate Cox regression, while regression on the subdistribution hazard has been obtained by Fine & Grey's method and Approximate Fine & Grey regression. Due to few cancer occurrences in relation to persons at risk, regression on the cause specific hazard and the subdistribution hazard gives approximately the same result. Hence, regression on the cause specific hazard is used as a reference method for a complete medical analysis.

There are many weaknesses with this analysis, such as for example incomplete data. Participants with missing values are taken out of the analysis when the explanatory variable is included in the model, this makes the analysis less reliable compared to analysis with complete data. The problem of missing data also makes it difficult to find an optimal model to describe the risk of breast, uterine and ovarian cancer in the presence of the competing events. It is natural to believe that the risk of cancer is heritable, factors such as family relationships are not taken into account in this analysis. It is also important to mention that the covariate values most likely vary throughout the 50 year study. Many women will gain or lose weight with time, this model does not consider any such circumstances. However, data on height and weight was assigned to the study some years after the start of follow up, the error of misclassified weight is therefore assumed to be less than it would have been with initial weight, as it is closer to the terminal date for most participants.

For a further analysis, it is possible to cover some of the shortcomings, such as

missing values. It is also possible to include more competing events, as for example other cancer types.

Within a certain framework, when there are explanatory variables in the model, identification of the underlying dependence structure between the individual component in latent failure time representation is possible, see Heckman and Honore [14]. For a further analysis, it would have been interesting to investigate this further.

Briefly, the main medical findings of this thesis are:

• Several births affect the risk of getting breast, uterine or ovarian cancer:
Each additional birth decreases the risk of getting breast, uterine or ovarian cancer by 10 %, 10 % and 16%, respectively.

• Age at first birth affect the risk of getting breast or uterine cancer:
Women with their first birth after the age of 30 years have approximately 41 % reduced risk of getting uterine cancer compared to women who gives the first birth before the age of 20. The opposite effect can be seen for breast cancer risk, with 45% increased risk of breast cancer for women with the first birth after the age of 30 compared to first birth before the age of 20.

• Age at last birth affect the risk of uterine and breast cancer:
Early last birth is protective against breast cancer compared to late last birth. The opposite result applies to uterine cancer, where a late last birth seems protective.

• Age at menarche affects the risk of getting uterine and breast cancer:
Each year increase in age at menarche decreases the risk of uterine and breast cancer with approximately 12% and 5%, respectively. Ovarian cancer risk is not affected by age at menarche.

• Age at menopause affects the risk of getting breast and uterine cancer:
High age at menopause increases the risk of breast and uterine cancer, with approximately 3% and 6 % for each additional year, respectively.

• Obesity affects the risk of getting breast, uterine and ovarian cancer:
Obesity is associated with increased risk of ovarian cancer for postmenopausal women. The risk of getting uterine and breast cancer increases with approximately 7% and 2% for each additional unit increase in BMI, respectively.

References

[1] Cancer registry of norway. http://kreftregisteret.no/.

[2] The comprehensive r archive network. http://cran.r-project.org.

[3] World health organization. http://www.who.int/en/.

[4] Cancer research uk. http://www.cancerresearchuk.org/.

[5] The norwegian cancer society. http://www.kreftforeningen.no/.

[6] T. Bjørge, A. Engeland, S. Tretli, and E. Weiderpass. Body size in relation to cancer of the uterine corpus in 1 million norwegian women. *International Journal of Cancer*, 120:378–383, 2006.

[7] M.J. Crowder. *Classical Competing Risks*. Chapman & Hall/CRC: Boca Raton, 1 edition, 2001.

[8] L. Dossus et al. Reproductive risk factors and endometrial cancer: the european prospective investigation into cancer and nutrition. *International Journal of Cancer*, 127:442–451, 2009.

[9] J.P. Fine and R.J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446): 496–509, 1999.

[10] R.B. Geskus. Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, 67: 39–49, 2011.

[11] R.J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141–1154, 1988.

[12] L. Grude. Competing risks. Specialization Project TMA4500 Industrial Mathematics, NTNU, 2010.

[13] S.E. Hankinson et al. A prospective study of reproductive factors and risk of epithelial ovarian cancer. *Cancer*, 76(2):284–290, 1995.

[14] J.J. Heckman and B.E. Honore. The identifiability of the competing risks model. *Biometrika*, 76:325–330, 1989.

[15] B.S. Krogstad. Risikofaktorer for brystkreft, en oppfølgingsstudie av 38178 kvinner i nord-trøndelag. Master's thesis, Industrial Mathematics, NTNU, 1998.

[16] G. Kvåle. *Reproductive factors and risk of cancer of the breast and genital organs, A prospective study of Norwegian women*. PhD thesis, Department of hygiene and social medicine, University of Bergen, 1989.

[17] G. Kvåle and I. Heuch. A prospective study of reproductive factors and breast cancer, II. Age at first and last birth. *American Journal of Epidemiology*, 126 (5):842–850, 1987.

[18] G. Kvåle and I. Heuch. Lactation and cancer risk: is there a relation specific to breast cancer? *Journal of Epidemiology and Community Health*, 42(1):30–37, 1987.

[19] G. Kvåle and I. Heuch. Menstrual factors and breast cancer risk. *Cancer*, 62 (8):1625–1631, 1988.

[20] G. Kvåle, I. Heuch, and G.E. Eide. A prospective study of reproductive factors and breast cancer, I. Parity. *American Journal of Epidemiology*, 126(5):831–841, 1987.

[21] G. Kvåle, I. Heuch, S. Nilssen, and V. Beral. Reproductive factors and risk of ovarian cancer: a prospective study. *International Journal Of Cancer*, 42(2): 246–251, 1988.

[22] G. Kvåle, I. Heuch, and G. Ursin. Reproductive factors and risk of cancer of the uterine corpus: A prospective study. *Cancer research*, 48:6217–6221, 1988.

[23] J.P. Lawless. *Statistical models and methods for lifetime data*. Wiley, 2 edition, 2003.

[24] B.H. Lindqvist. Competing risks, in: Encyclopedia of statistics in quality and reliability. *John Wiley & Sons Ltd*, pages 335–341, 2008.

[25] B.H. Lindqvist. Competing risks in a health survey. Unpublished lecture notes. Norwegian University of Science and Technology, Trondheim, Norway, 2011.

[26] C.P. McPherson, T.A. Sellers, J.D. Potter, R.M. Bostick, and A.R. Folsom. Reproductive factors and risk of endometrial cancer the iowa women's health study. *American Journal of Epidemiology*, 143(12):1195–1202, 1996.

[27] K. McPherson, C.M. Steel, and J.M. Dixon. Abc of breast diseases, breast cancer-epidemiology, risk factors, and genetics. *British Medical Journal*, 321: 625–628, 2000.

[28] A.V. Peterson. Bounds for a joint distribution function with fixed sub-distribution functions: applications to competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 73(1):11–13, 1976.

[29] M. Pintilie. *Competing Risks: A Practical Perspective (Statistics in Practice)*. Wiley, 1 edition, 2006.

[30] R.L. Prentice, J.D. Kalbfleisch, A.V. Peterson, N. Flournoy, V.T Farewell, and N.E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978.

[31] J.M. Ramon et al. Age at first full-term pregnancy, lactation and parity and risk of breast cancer: A case-control study in spain. *European Journal of Epidemiology*, 12(5):449–453, 1996.

[32] G.A. Satten and S. Datta. The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3): 207–210, 2001.

[33] L. Titus-Ernstoff et al. Menstrual and reproductive factors in relation to ovarian cancer risk. *British Journal Of Cancer*, 84(5):714–721, 2001.

[34] A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.

[35] M. Zheng and J.P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.