UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 153

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 153

EFFICIENT DESIGN AND MODELING STRATEGIES FOR FOLLOW-UP STUDIES WITH TIME-VARYING COVARIATES

JAAKKO REINIKAINEN

JYVÄSKYLÄ
2015

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 153

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 153

EFFICIENT DESIGN AND MODELING STRATEGIES FOR FOLLOW-UP STUDIES WITH TIME-VARYING COVARIATES

JAAKKO REINIKAINEN

To be presented, with the permission of the Faculty of Mathematics and Science
of the University of Jyväskylä, for public criticism in Auditorium MaA211
on November 20th, 2015, at 13 o'clock.

JYVÄSKYLÄ
2015

Editor: Pekka Koskela
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI–40014 University of Jyväskylä
Finland

ISBN 978-951-39-6429-0 (PDF)

ISBN 978-951-39-6315-6
ISSN 1457-8905

University Printing House
Jyväskylä 2015

Abstract

Epidemiological studies can often be designed in several ways, some of which may be more optimal than others. Possible designs may differ in the required resources or the ability to provide reliable answers to the questions under study. In addition, once the data are collected, the selected modeling approach may affect how efficiently the data are utilized.

The purpose of this dissertation is to investigate efficient designs and analysis methods in follow-up studies with longitudinal measurements. A key question is how to select optimally a subcohort for a new longitudinal covariate measurement if we cannot afford to measure the entire cohort. Another key question we consider is how to determine the reasonable number of longitudinal measurements. Different ways to utilize longitudinal covariate measurements in modeling cardiovascular disease (CVD) mortality are also studied.

Follow-up data are modeled using parametric or semiparametric proportional hazards models. Subcohort selections are carried out using optimality criteria initially developed for optimal experimental design. Measures of model discrimination are applied to plan the number of longitudinal measurements. The topics are studied using simulations and the East–West data, which are Finnish part of an international follow-up study in the field of cardiovascular epidemiology, the Seven Countries Study.

This work demonstrates that the cost-efficiency of follow-up designs can be improved by careful planning. The proposed method for selecting optimal subcohorts is shown to outperform simple random sampling and it is demonstrated how the number of longitudinal measurements can be determined using simulated data and data from previous similar studies. The results also indicate that individual-level changes and cumulative averages of classical risk factors are good predictors of CVD mortality.

Keywords: follow-up study, time-varying covariates, longitudinal measurements, optimal design, data collection, risk prediction, cardiovascular disease mortality

Tiivistelmä

Epidemiologiset tutkimukset voidaan usein toteuttaa monella eri tavalla, joista toiset saattavat olla optimaalisempia kuin toiset. Mahdolliset tutkimusasetelmat voivat erota niiden edellyttämissä resursseissa tai kyvyssä tarjota luotettavia vastauksia tutkimuskysymyksiin. Lisäksi valittu menetelmä aineiston mallintamiseen voi vaikuttaa siihen, kuinka tehokkaasti kerättyä aineistoa pystytään hyödyntämään.

Tämän väitöskirjan tavoitteena on tutkia tehokkaita tutkimusasetelmia ja analyysimenetelmiä pitkittäismittauksia sisältävissä seurantatutkimuksissa. Keskeisenä kysymyksenä on, miten alkuperäisestä kohortista valitaan optimaalisesti osajoukko uuteen kovariaattien pitkittäismittaukseen, jos ei ole varaa mitata koko kohorttia. Toinen käsiteltävä kysymys on, miten valitaan riittävä määrä pitkittäismittauksia. Työssä tarkastellaan myös erilaisia tapoja hyödyntää kovariaattien pitkittäismittauksia sydän- ja verisuonitautikuolleisuuden mallinnuksessa.

Seuranta-aineistoa mallinnetaan parametrisillä tai semiparametrisillä suhteellisen vaaran malleilla. Osakohortin valinnassa käytetään alunperin optimaalisiin kokeellisiin tutkimusasetelmiin kehitettyjä optimaalisuuskriteerejä. Pitkittäismittausten lukumäärän suunnittelussa sovelletaan mallin erottelukyvyn mittareita. Tutkimuksessa käytetään simulointikokeita ja Itä–Länsi-aineistoa, joka on suomalainen osa kansainvälistä sydän- ja verisuonitauteihin keskittyvää Seitsemän maan tutkimus -seurantatutkimusta.

Tuloksista nähdään, että seurantatutkimusten kustannustehokkuutta voidaan parantaa huolellisella suunnittelulla. Esitetty menetelmä osakohortin optimaaliseen valintaan osoittautuu yksinkertaista satunnaisotantaa paremmaksi. Tutkimus havainnollistaa, miten simulointeja ja aineistoja aiemmista samankaltaisista tutkimuksista voidaan hyödyntää pitkittäismittausten lukumäärän valinnassa. Tulokset osoittavat myös, että klassisten riskitekijöiden yksilötason muutokset ja kumulatiiviset keskiarvot ovat hyviä selittäjiä sydän- ja verisuonitautikuolleisuudelle.

Avainsanat: seurantatutkimus, aikariippuvat kovariaatit, pitkittäismittaukset, optimaalinen tutkimusasetelma, aineiston keruu, riskin ennustaminen, sydän- ja verisuonitautikuolleisuus

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Juha Karvanen. It has been a pleasure to work with him, as he has been full of new ideas and has always had time for me. During the last three years I have learned a lot about statistics and research in general from Juha.

I have had a great opportunity to collaborate with the National Institute for Health and Welfare. Especially, I wish to thank my co-authors Dr. Hanna Tolonen and Prof. Tiina Laatikainen, whose expertise in epidemiology was invaluable in this work.

I am grateful to Prof. Kari Auranen and Prof. Richard J. Cook for reviewing this thesis and providing constructive comments. I would also like to give my thanks to all the colleagues at the Department of Mathematics and Statistics, University of Jyväskylä, for many interesting discussions and for making these years enjoyable. Tuula Blåfield is greatly acknowledged for proof-reading the introductory part of the thesis.

This work was funded by Emil Aaltonen Foundation and Department of Mathematics and Statistics, University of Jyväskylä. I am thankful for the grants that made this research possible. I also wish to thank CSC – IT Center for Science Ltd. for providing computing resources.

I am extremely grateful to my family members who have supported me throughout my life. In addition, I have several friends who deserve to be acknowledged for making my free time fun and relaxing. Finally, my special thanks go to Katja, whose support and patience have given me strength and confidence during this project.

Nurmijärvi, October 2015

Jaakko Reinikainen

List of original publications

This thesis consists of an introductory part and the following publications, which are referred to in the text by their Roman numerals.

I Reinikainen, J., Karvanen, J. and Tolonen, H. (2014). Optimal selection of individuals for repeated covariate measurements in follow-up studies. *Statistical Methods in Medical Research*, DOI: 10.1177/0962280214523952.

II Reinikainen, J. and Karvanen, J. (2015). Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies. *Submitted*.

III Reinikainen, J., Karvanen, J. and Tolonen, H. (2015). How many longitudinal covariate measurements are needed for risk prediction?. *Journal of Clinical Epidemiology*, DOI: 10.1016/j.jclinepi.2015.06.022.

IV Reinikainen, J., Laatikainen, T., Karvanen, J. and Tolonen, H. (2015). Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland. *International Journal of Epidemiology*, 44(1): 108-116.

Author's contribution

The author of this dissertation has performed all the analyses and simulations and had the main responsibility in writing the joint papers I–IV. The author has also derived the selection procedures used in Articles I and II. The research questions and statistical models have been formulated in collaboration with the co-authors. In addition, the co-authors have contributed in writing minor parts of the papers and interpreting the results, especially the epidemiological content in Article IV.

Contents

Chapter 1

Introduction

The efficient allocation of available resources is a common problem in the planning of data collection. The precision of estimates can be improved by collecting more data, but this usually increases the costs of the study, which has motivated researchers to consider the cost-efficiency of different study designs. The study is often designed either to maximize the precision for a fixed budget or to achieve the desired precision with costs as low as possible.

In epidemiological research, the aim is usually to study the effects of different risk factors on disease conditions or on lifetime in defined populations (Carneiro and Howard, 2011; Krickeberg et al., 2012). These kind of studies require the follow-up of individuals and the measuring of potential risk factors. When the interest lies in the survival of individuals, a prospective cohort study (Euser et al., 2009) may be appropriate. Assume, for example, that we would like to investigate how certain risk factors, called covariates in what follows, are related to lifetimes of individuals in some population. The study could be conducted by carrying out measurements of the covariates for a cohort of individuals and by following the cohort for mortality over a predetermined period of time.

Many covariates are time-varying, which means that their values do not remain constant in time (Dekker et al., 2008). For example, blood pressure, cholesterol and body mass index are time-varying, whereas time and place of birth and sex are examples of time-fixed or time-invariant variables. Measuring time-varying covariates repeatedly brings more information on them and may lead to a higher precision in the estimation of the covariate effects. There are also other ways how the follow-up design could be improved. Increasing the number of individuals in the cohort, lengthening the follow-up period or measuring additional variables could increase the precision of the estimates. However, all these actions would require more resources, and so they provide opportunities for study design optimization.

The research problems of this thesis are based on epidemiological follow-up studies, where time-varying covariates are present. The aim is to give answers to the questions outlined below:

(A) Assume that longitudinal measurements are carried out for the covariates, but we

cannot afford to measure the entire cohort in re-measurements. How to select the subcohort to be re-measured in order to obtain precise estimates of the covariate effects in a survival model?

(B) How to choose the number of longitudinal covariate measurements cost-efficiently, when the survival model will be used for risk prediction?

(C) How to utilize the longitudinal covariate measurements in the modeling of cardiovascular disease (CVD) mortality?

There are many other design problems that fall out of the scope of this thesis, such as: the optimal combination of the number of individuals in the cohort and the number of longitudinal measurements for each individual, finding the solution to Question (A) when the aim is to obtain precise risk predictions, or answering Question (B) when the model parameters rather than risk predictions are of interest. Moreover, we do not consider questions related to the target population, the sampling frame, the size of the original cohort, or the covariates to be measured.

Question (A) is studied in Articles I and II. First, Article I considers the subcohort selection with two measurement times and one covariate and Article II extends the results to cases where several covariates and measurement points are allowed. Article III investigates Question (B) by considering how measures of model discrimination (Steyerberg et al., 2010) can be used to help the decision about the reasonable number of longitudinal measurements. We propose measuring the improvement in model discrimination between models that use different amounts of longitudinal information. Questions (A) and (B) are about designing the study cost-efficiently, whereas Question (C) concerns the analysis of collected data in a real follow-up study. This is further studied in Article IV.

Figure 1.1 is a simplified illustration of the follow-up design we are considering and exemplifies in which phases of a study the results and methods of Articles I–IV could be applied. The figure presents all three research questions together although they are addressed separately in Articles I–IV. In Articles I and II, it is assumed that the entire cohort is measured at the baseline. The design is then constructed sequentially by selecting a subcohort optimally just before each re-measurement from those individuals still alive. All previously collected data can be utilized in the selections. In the figure, a subcohort consisting of individuals 1 and 2 is selected for the second measurement. Individual 3 could not have been selected, because he/she has already experienced the event of interest. The subcohort for the mth measurement consists of individuals 2 and 5. At the end of the follow-up, individuals 2 and 4 have not experienced the event, so their lifetimes are said to be censored (Collett, 2003). Article III considers the problem of choosing the number of measurement points for both ongoing and completely new studies, so it could have been placed as well before the baseline measurement in Figure 1.1.

Follow-up studies create survival data, which we model using parametric or semi-parametric proportional hazards models (Kalbfleisch and Prentice, 2002; Aalen et al.,

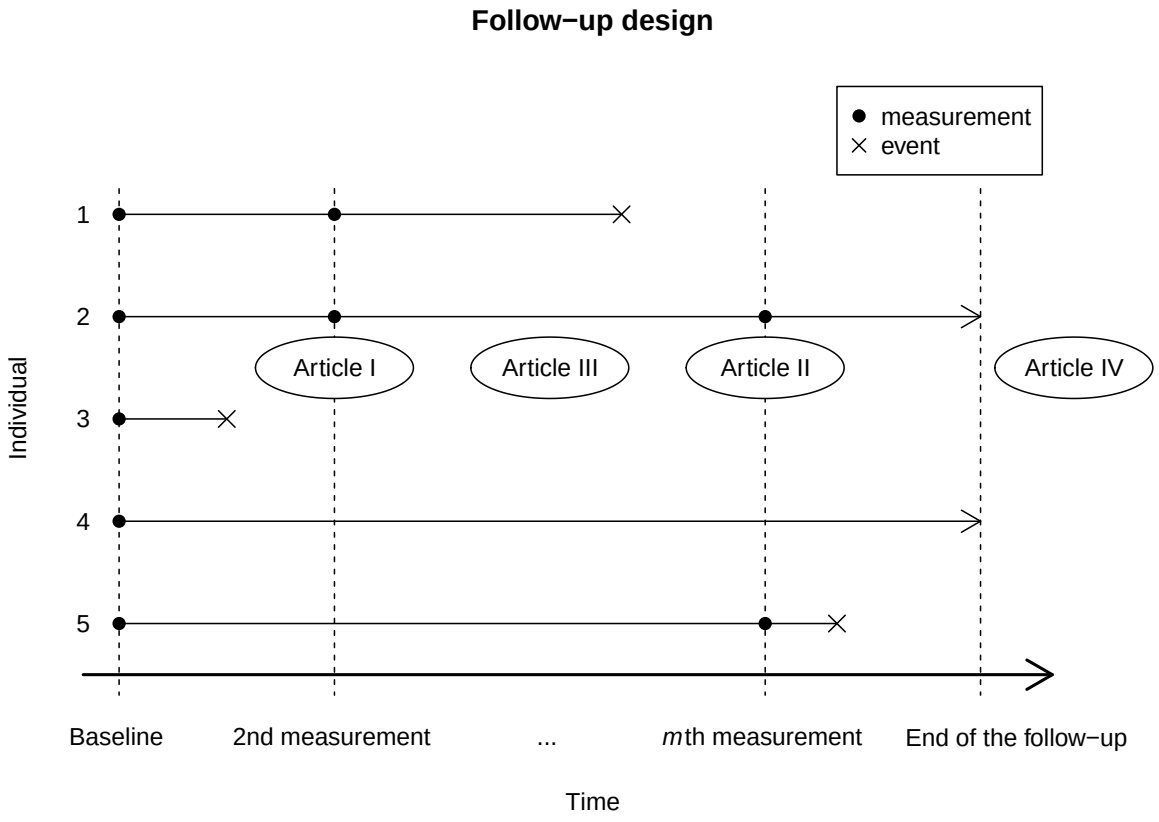

Figure 1.1: A simplified example of a follow-up design, where five individuals are followed and the covariates in subsets of two individuals are re-measured. The figure also exemplifies in which phases of a study the results and methods of Articles I–IV could be applied.

2008). In Articles I and II, only a subset of the cohort is measured in re-measurements, which leads to a large amount of missing data. We use a likelihood-based approach with numerical integration, multiple imputation and Bayesian data augmentation (Little and Rubin, 2002) to handle missing data in the analyses. The subcohort selections are carried out using optimality criteria based on the Fisher information matrix (Pukelsheim, 1993; Atkinson et al., 2007). These criteria were originally developed for the design of experiments, but have later been applied also in observational studies (Karvanen et al., 2009; Mehtälä et al., 2011).

In Article III, the study planning is based on model performance comparisons. The performance is evaluated using different measures of model discrimination: the area under the receiver operating characteristic curve (Hanley and McNeil, 1982), the net reclassification improvement index (Pencina et al., 2011), the integrated discrimination improvement index (Pencina et al., 2008) and the net benefit (Vickers and Elkin, 2006). Measures of model discrimination are also used in Article IV to compare different ways

of using longitudinal measurements in the analysis of CVD mortality. In Article IV, we derive new time-varying covariates to study the use of individual-level cumulative averages and changes in the classical risk factors of CVD. The cumulative averages are calculated as averages of the most recent and all previous measurements and the changes as differences of the latest two measurements.

In Articles I–IV, the research questions are approached using data from the Finnish cohorts of the Seven Countries Study, and in Articles I–III also simulation studies are employed. The Seven Countries Study was initiated in the late 1950s as one of the first international follow-up studies in the field of cardiovascular epidemiology. Its objective was to investigate the development of CVD and related risk factors in different countries (Keys, 1970; Menotti and Puddu, 2013). The study had two cohorts in Finland, one in Eastern and another in South–Western Finland, from which comes the name East–West study (Karvonen et al., 1966). The cohorts consisted of all men born between 1900 and 1919 in two geographically defined areas, 1711 men in total. A large set of variables thought to be possibly related to CVD was measured in 1959, 1964, 1969, 1974, 1984, 1989, 1994 and 1999. Data on individuals' lifetimes and causes of death are available up to the end of 2011.

The theoretical framework of this thesis is introduced in Chapters 2 and 3. Chapter 2 considers the analysis of survival data with some special issues, and Chapter 3 reviews the theory of optimal experimental design and discusses its application in observational studies. The theory in these chapters is presented mainly from the epidemiological point of view with emphasis on the methods applied in Articles I–IV. The research contribution of this thesis is summarized in Chapter 4. Finally, Chapter 5 summarizes the results and discusses their implications and possible topics for future research.

Chapter 2

Survival models

Survival data arise in many fields where the time until the occurrence of an event is of interest. These times may be called survival times, lifetimes, failure times, or event times. The analysis of survival data has numerous applications in medical sciences, where survival time can be defined as the time until death or the occurrence of some disease, for instance. Other examples of survival times include the time until the failure of a machine, time taken to complete a task in a psychological experiment and time from the beginning of studies until graduation. Often the interest in survival analysis is in the dependence of survival times on explanatory variables. This chapter first introduces some models for survival data and then considers a few special issues in survival analysis.

2.1 Parametric models

There are some special features in survival data, which set requirements for modeling. Survival times are nonnegative and have often highly skewed distributions. However, the main feature that characterizes survival data is the presence of censored observations. If an individual has not experienced the event of interest at the end of the follow-up or has become lost during the follow-up, we know only that the survival time is greater than some constant c. This is called right censoring and c is a censored survival time. See, e.g., Collett (2003) for discussion on other types of censoring.

We denote the random variable of survival time by T_i and the observed survival time by t_i for individual i. In practice, a pair (t_i, δ_i) is observed, where δ_i is a status indicator, which is 1 if the actual survival time is observed and 0 if it is censored. When $\delta_i = 0$, t_i is the censoring time. Next, we define two basic functions used in survival analysis, namely the survival function $S(t)$ and the hazard function $\lambda(t)$. These functions and their basic results form the basis for survival analysis and are presented, for example, in textbooks by Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), and Aalen et al. (2008)

Suppose that the continuous random variable T has a probability density function

$f(t)$. Then, the cumulative distribution function can be expressed as

$$F(t) = P(T < t) = \int_0^t f(u)du.$$

This is the probability that the survival time is less than t, whereas the survival function gives the probability that the survival time is greater than or equal to t:

$$S(t) = P(T \geq t) = 1 - F(t).$$

The hazard function is defined as the instantaneous rate at which the event occurs at time t conditional on the individual has survived until time t

$$\lambda(t) = \lim_{\Delta \to 0+} \frac{P(t \leq T < t + \Delta | T \geq t)}{\Delta}.$$

The hazard can be interpreted as an instantaneous risk of the event.

From these definitions we obtain some useful results. The hazard function can be written as

$$\lambda(t) = \frac{f(t)}{S(t)}$$

and the survival function as

$$S(t) = e^{-\Lambda(t)},$$

where

$$\Lambda(t) = \int_0^t \lambda(u)du$$

is the cumulative hazard function. These results allow us to derive the density function, survival function and hazard function once one of them is known.

There are many different models developed for survival analysis to assess the relationship between survival times and covariates. Proportional hazards models play a central role in practical survival analysis and are therefore mainly considered here. Two other classes of survival models, namely accelerated failure time models and proportional odds models, are treated briefly.

Assume that we have covariates $\boldsymbol{x} = (x_1, \ldots, x_H)^T$, upon which survival times may depend. The covariate vector $\boldsymbol{x}$ may include continuous or binary variables, interactions between them or quadratic terms of original variables, for instance. Proportional hazards models (Collett, 2003), sometimes called relative risk models (Kalbfleisch and Prentice, 2002), are of the form

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}^T\boldsymbol{\beta}}, \tag{2.1}$$

where $\lambda_0(t)$ is a baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_H)^T$ are the regression coefficients. In this model, the covariates have multiplicative effects on the hazard. The baseline hazard function is left unspecified in the semiparametric Cox proportional hazards model (Cox, 1972), which is considered in Section 2.2. For example, if we

assume the Weibull or Gompertz distribution (Bender et al., 2005) for the survival times, we obtain a parametric proportional hazards model.

If we use Model 2.1, the density function is

$$f(t|\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}^T\boldsymbol{\beta}} \exp\left[-e^{\boldsymbol{x}^T\boldsymbol{\beta}}\Lambda_0(t)\right]$$

and the survival function is

$$S(t|\boldsymbol{x}) = \exp\left[-e^{\boldsymbol{x}^T\boldsymbol{\beta}}\Lambda_0(t)\right].$$

For the Weibull distribution, parameterized as

$$f(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \exp\left[-\left(\frac{t}{b}\right)^a\right],$$

where $a > 0$ is a shape parameter and $b > 0$ is a scale parameter, the hazard function, survival function, and density function are written as

$$\begin{aligned}
\lambda_{\text{Weib}}(t|\boldsymbol{x}) &= \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} e^{\boldsymbol{x}^T\boldsymbol{\beta}}, \\
S_{\text{Weib}}(t|\boldsymbol{x}) &= \exp\left[-e^{\boldsymbol{x}^T\boldsymbol{\beta}}\left(\frac{t}{b}\right)^a\right] \text{ and} \\
f_{\text{Weib}}(t|\boldsymbol{x}) &= \lambda_{\text{Weib}}(t|\boldsymbol{x})S_{\text{Weib}}(t|\boldsymbol{x}).
\end{aligned}$$

Accelerated failure time (AFT) models are log-linear models for the event time T. The model specifies that

$$\log T = \boldsymbol{x}^T\boldsymbol{\beta} + W, \tag{2.2}$$

where W is an error variable with density f_W. Model (2.2) can also be written as

$$T = e^{\boldsymbol{x}^T\boldsymbol{\beta}}e^W.$$

From this form we see that the covariates have multiplicative effects on the event times rather than on the hazard. Assume that e^W has the hazard function $\lambda_0(t)$. It then follows that

$$\lambda(t|\boldsymbol{x}) = e^{-\boldsymbol{x}^T\boldsymbol{\beta}}\lambda_0(te^{-\boldsymbol{x}^T\boldsymbol{\beta}}).$$

The common choices for the distributions of survival times in Model (2.2) include the Weibull, log-logistic and log-normal distributions. These lead to the Gumbel, logistic, and normal distributions for W, respectively. The only AFT models that are also proportional hazards models are those which assume the exponential or Weibull distribution for the survival times (Cox and Oakes, 1984).

The third class of survival models presented here is the proportional odds models (Collett, 2003). In these models the odds of an individual surviving beyond time t are modeled as

$$\frac{S(t)}{1-S(t)} = e^{\boldsymbol{x}^T\boldsymbol{\beta}}\frac{S_0(t)}{1-S_0(t)}. \tag{2.3}$$

In the proportional odds model, the covariates have multiplicative effects on the odds of survival beyond t. A common choice for the distribution of survival times in Model (2.3) is the log-logistic distribution. Actually, this is the only distribution which has both the proportional odds property and the accelerated failure time property (Smithson and Merkle, 2014).

The parametric models introduced above can be fitted using the maximum likelihood method. The likelihood function for data with n individuals, in the presence of right censored survival times, is of the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{x}_i)^{\delta_i} S(t_i|\boldsymbol{x}_i)^{1-\delta_i},$$

where $\boldsymbol{\theta}$ is a vector of the model parameters. In some applications, alternative definitions for the time-origin of survival times may be possible. In epidemiology, the time-scale can be, for example, age, calendar time, time-on-study, or time since diagnosis. Kom et al. (1997) and Thiébaut and Bénichou (2004) recommend to use age as the time-scale. If the survival times are not observed from the time origin, this must be taken into account in modeling by using truncated distributions. When age is used as the time-scale and the cohort members are 50 years old at the baseline of the study, for instance, the distribution of survival times is left-truncated at 50 years. Let t_0 be the truncation time. Then, the likelihood function becomes

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{f(t_i|\boldsymbol{x}_i)}{S(t_0|\boldsymbol{x}_i)} \right)^{\delta_i} \left(\frac{S(t_i|\boldsymbol{x}_i)}{S(t_0|\boldsymbol{x}_i)} \right)^{1-\delta_i}.$$

2.2 Semiparametric models

In the proportional hazards model (2.1), the baseline hazard function can be left unspecified (Cox, 1972). This leads to the semiparametric Cox proportional hazards model (briefly the Cox model), which is widely used due to its flexibility. The method for estimating the parameters $\boldsymbol{\beta}$ is called partial likelihood (Cox, 1972, 1975).

Assume first that there are no ties in the data, i.e., only one individual dies at each death time. The partial likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} e^{\boldsymbol{x}_l^T \boldsymbol{\beta}}} \right)^{\delta_i}, \tag{2.4}$$

where $R(t_i)$ is a so-called risk set. The set $R(t_i)$ includes individuals who are known to be alive just before the time t_i. Likelihood (2.4) is called partial likelihood, because it is not a full likelihood as it does not use the actual survival times directly, but depends only on the ranking of these times.

In practice, survival data often include tied observations. In order to estimate the $\boldsymbol{\beta}$ parameters in the presence of ties, the partial likelihood function (2.4) has to be

modified. Let $\boldsymbol{s}_j$ be a sum of the covariate vectors for the individuals dying at the jth death time from the ordered death times, $t_{(j)}$, $j = 1, \ldots, r$, where r is the number of different death times. The number of deaths at $t_{(j)}$ is denoted by d_j. If there are not very many ties at any death time, an adequate approximation (Breslow, 1974) is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{e^{\boldsymbol{s}_j^T \boldsymbol{\beta}}}{\left(\sum_{l \in R(t_{(j)})} e^{\boldsymbol{x}_l^T \boldsymbol{\beta}} \right)^{d_j}}.$$

Other approximations for the partial likelihood have been proposed, e.g., by Efron (1977) and Kalbfleisch and Prentice (2002).

Semiparametric AFT models are not as commonly used as Cox models because of computational difficulties. However, semiparametric AFT models can be estimated using rank-based weighted generalized estimating equations (Chiou et al., 2014).

2.3 Time-varying covariates

In the previous sections, only the baseline characteristics of the study subjects were used as covariates in survival models. Some variables, however, change over time during the study, and taking this into account in modeling may improve the estimation of the covariate effects. Such covariates are called time-varying or time-dependent.

Time-varying covariates are often classified as either to be internal or external (Collett, 2003). A covariate is external, if its future path is not affected by the occurence of the event of interest. If a covariate is not external, it is internal. Examples of internal variables include blood pressure, cholesterol and smoking status. Usually, information on internal variables is obtained by measuring them, and it requires the measured individual being alive. External variables do not necessarily require an individual being alive for their existence. Policy changes, the dose of a drug, or the treatment group in which a patient is assigned are examples of external variables.

The two main approaches for modeling survival data with time-varying covariates are time-dependent Cox models (Therneau and Grambsch, 2000) and the joint modeling of longitudinal and survival data (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Rizopoulos, 2012). We denote the vector of covariate values of the individual i at time t by $\boldsymbol{x}_i(t)$, which may include both fixed and time-varying covariates. When the proportional hazards model (2.1) is extended to incorporate the time-varying covariates, it becomes

$$\lambda(t|\boldsymbol{x}_i(t)) = \lambda_0(t) e^{\boldsymbol{x}_i(t)^T \boldsymbol{\beta}}. \tag{2.5}$$

Note that as the values $\boldsymbol{x}_i(t)$ may vary in time, the relative hazard $\lambda(t|\boldsymbol{x}_i(t))/\lambda_0(t)$ does not remain constant over time. This means that the model (2.5) is actually not a proportional hazards model.

For external covariates whose values are known at every time point, the time-dependent Cox model (2.5) is appropriate. When we use this model, the partial likeli-

hood (2.4) becomes

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{x}_i(t_i)^T \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} e^{\boldsymbol{x}_l(t_i)^T \boldsymbol{\beta}}} \right)^{\delta_i}.$$

Let us assume that the covariates change their values only at certain time points, i.e., the covariate processes are step functions. Now, we can write the likelihood contribution of the individual i in a parametric model as

$$L_i(\boldsymbol{\theta}) = \prod_{m=0}^{M_i} \left(\frac{f(t_{m+1,i}|\boldsymbol{x}_{m,i})}{S(t_{m,i}|\boldsymbol{x}_{m,i})} \right)^{\delta_{m+1,i}} \left(\frac{S(t_{m+1,i}|\boldsymbol{x}_{m,i})}{S(t_{m,i}|\boldsymbol{x}_{m,i})} \right)^{1-\delta_{m+1,i}}, \tag{2.6}$$

where M_i is the number of time points $t_{m,i}$ where the covariate values $\boldsymbol{x}_{m,i}$ change and status indicators $\delta_{1,i}, \ldots, \delta_{M_i,i}$ are zeros, and $\delta_{M_i+1,i}$ may be zero or one. Note that the denominators in the likelihood (2.6) are needed because we have to deal with truncated distributions as the time scale is divided according to the intervals where the covariates remain constant.

The estimates of the time-dependent Cox model (2.5) may be biased, when internal time-varying covariates are observed in discrete time and measured with error (Prentice, 1982; Bycott and Taylor, 1998). In this case, the joint modeling of longitudinal and survival data is an appropriate approach (Rizopoulos, 2012).

A central idea underlying the so called joint models is that the survival time data and longitudinal covariate data are assumed to depend on the common set of latent random effects (Tsiatis and Davidian, 2004). Originally, two-stage methods were proposed to fit joint models (Tsiatis et al., 1995). The idea is to estimate first the parameters of the longitudinal model and then to use the estimated model as a covariate in the survival model. However, this kind of approach may lead to biased estimates (Dafni and Tsiatis, 1998). For this reason, nowadays the preferred method for fitting joint models is the maximum likelihood method based on the joint distribution of the observed survival and longitudinal data (Henderson et al., 2000; Hsieh et al., 2006).

Despite the major developments in joint modeling methodology (Brown and Ibrahim, 2003; Rizopoulos et al., 2009; Rizopoulos, 2011), it has been demonstrated that the time-dependent Cox model should still be considered a candidate when selecting the most suitable method for analysis (Hanson et al., 2011). In addition, when there are multiple time-varying covariates, problems arise in fitting joint models due to increased computational complexity (Rizopoulos, 2012), and therefore some specialized methods have been developed for multivariate joint modeling (Brown et al., 2005; Proust-Lima et al., 2009; Rizopoulos and Ghosh, 2011). Bayesian methods have been favored in the estimation of complex joint models (Gould et al., 2015).

Whatever is the selected approach for modeling survival data with longitudinal covariate measurements, the analyst should also consider whether to use the original measurements or some derived variables. These derived variables may, for instance, be averages of the most recent and all the previous measurements (Wilson et al., 1997), differences of the latest two measurements (Farchi et al., 1981; Sesso et al., 2000),

standard deviation of the measurements (Muntner et al., 2011), maximum value reached (Rothwell et al., 2010), or lagged observations (Hanson et al., 2011). The averages and differences of longitudinal measurements were used in the modeling of CVD mortality in Article IV, whose contribution is summarized in Section 4.4.

2.4 Missing data

Missing data are encountered in follow-up studies for various reasons. An individual may refuse to participate or is lost, for example, due to moving to another area. One important type of missingness is data missing by design, which arises, for example, if some measurements are carried out for only a subset of individuals due to budget limitations. All these may lead to missingness in the covariates or in the response variable. Removing individuals with missing information from the data set might lead to loss of power and biased estimates, and thus several methods have been developed to handle missing data. Three approaches for handling missing data will be briefly introduced: likelihood-based approach, multiple imputation and Bayesian data augmentation. These are applied in Articles I and II to handle data missing by design.

Rubin (1976) defined the following classes of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MCAR and MAR assumptions, the missing data mechanism is usually said to be ignorable, which means that it is not necessary to specify a model for the mechanism of missingness. If data are assumed to be MNAR, the missingness is said to be informative and an additional model for missing data mechanism is required.

Let us assume that we have data $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$, which consist of the observed part $\boldsymbol{Y}_{obs}$ and the missing part $\boldsymbol{Y}_{mis}$. Missing data are assumed to be MAR or MCAR. In the likelihood-based approach, the inference is based on the likelihood

$$L(\boldsymbol{\theta}) = f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}) = \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}) d\boldsymbol{Y}_{mis}, \tag{2.7}$$

where the intergal is over the support of $\boldsymbol{Y}_{mis}$. Equation (2.7) specifies a likelihood of the model parameters $\boldsymbol{\theta}$. The missing data mechanism has been ignored here as we have assumed that the mechanism does not depend on the missing values. In a case where the variables including missing values are discrete, the integration would be replaced with summation. In practice, it may not be possible to calculate the integrals analytically, so numerical integration (Gautschi, 2012) can be applied. Alternative methods, such as the expectation–maximization (EM) algorithm (Dempster et al., 1977; Shen and Cook, 2013), can also be used to maximize the likelihood.

The idea of multiple imputation is to generate imputations for missing values m times ($m \geq 2$) using an imputation model, fit an analysis model with each imputed data set, and then combine the results to obtain the final estimates and standard errors (Rubin, 1987). Multiple imputation is usually preferred over single imputation techniques, because it takes into account the uncertainty about the parameters of the imputation model and also the uncertainty about the imputations.

Multiple imputation can be thought of as being an approximate Bayesian method, because the uncertainty of unknown parameters is expressed as posterior distributions. The posteriors of the imputation model parameters $\boldsymbol{\eta}$ are estimated using data without missing values. Then, realizations of these parameters $\boldsymbol{\eta}^*$ are drawn from the posteriors. Imputations are drawn from the conditional posterior distributions of missing values (the imputation model) given the realizations $\boldsymbol{\eta}^*$. This procedure is repeated m times. Then, the estimates for the parameters of the model of interest are obtained by fitting the analysis model for each imputed data set. Finally, the estimates and their standard errors are combined by applying Rubin's rules (Rubin, 1987), which take into account the within-imputation variance and between-imputation variance.

The most critical part in multiple imputation is the specification of the imputation model. Different methods have been proposed for this, including multivariate normal imputation (Lee and Carlin, 2010) and chained equations (White et al., 2011). White and Royston (2009) showed that when the analysis model is the proportional hazards model and the covariates include missing values, the status indicator δ_i and cumulative baseline hazard $\Lambda_0(t_i)$ should be used as predictors in the imputation model.

In the Bayesian approach, missing data values can be regarded as unknown parameters and treated in the analysis similarly to model parameters (Tanner and Wong, 1987). A prior distribution or model is specified for the missing values and their posteriors are calculated simultaneously with the posteriors for the model parameters, using, for example, MCMC methods (Lunn et al., 2012).

In addition to those presented above, there are also other methods for handling missing data, such as single imputation methods (Little and Rubin, 2002) and weighting techniques (Seaman and White, 2013). The choice of an appropriate method depends, e.g., on the research question, type and amount of missing data and the chosen paradigm.

2.5 Model performance

The goodness of a survival model can be evaluated using many different measures, which reflect different aspects of the model. If the model has been developed for risk prediction, its performance should be assessed primarily by measuring the predictive ability. The measures of predictive performance belong typically to one of the two main categories: model discrimination or model calibration (Cook, 2007). Discrimination means the model's ability to separate events and nonevents, whereas calibration quantifies how well the estimated risk and the actual risk agree. Here, we present some methods for evaluating the performance of survival models.

A well-known tool for investigating model discrimination is the receiver operating characteristic (ROC) curve (Metz, 1978; Fawcett, 2006). Assume that all individuals have either a positive (case) or negative (noncase) condition and a prediction model is developed to classify the individuals as positives or negatives. Let TP and FP denote the numbers of true positive and false positive classifications, respectively. The true positive rate, TPR, is the proportion of true positives among all individuals with a

Table 2.1: A contingency table showing how the true positive rate (TPR) and the false positive rate (FPR) are calculated.

		Condition	
		Positive	Negative
Classification by the model	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)
		True positive rate $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$	False positive rate $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$

positive condition and the false positive rate, FPR, is the proportion of false positives among all individuals with a negative condition, as displayed in Table 2.1. TPR is also known as sensitivity and (1− FPR) as specificity. The ROC curve is a plot of TPR against FPR at various risk thresholds and is illustrated in Figure 2.1.

The area under the ROC curve (AUC) is a discrimination metric, which summarizes the information of the ROC curve into a single value. The AUC can be interpreted as the probability that a randomly selected case has higher predicted risk than a random noncase (Hanley and McNeil, 1982). Two models can be compared by calculating the difference in AUCs.

More novel metrics for comparing the discrimination of two models include the net reclassification improvement (NRI) (Pencina et al., 2011) and the integrated discrimination improvement (IDI) (Pencina et al., 2008) indices. To estimate the NRI, we have to define risk categories and check how the new model reclassifies individuals compared to the old model. The estimate of the NRI is given by

$$\hat{\text{NRI}} = (\hat{p}_{\text{up, events}} - \hat{p}_{\text{down, events}}) - (\hat{p}_{\text{up, nonevents}} - \hat{p}_{\text{down, nonevents}}),$$

where $\hat{p}_{\text{up, events}} = (\#\text{ events moving up})/(\#\text{ events})$ and $\hat{p}_{\text{down, events}} = (\#\text{ events moving down})/(\#\text{ events})$. An event moving up (or down) means here a reclassification to a higher (or lower) risk category. Probabilities $\hat{p}_{\text{up, nonevents}}$ and $\hat{p}_{\text{down, nonevents}}$ are defined respectively. If there are no established risk categories, a category-less or continuous NRI can be used, where any upward or downward change in predicted probabilities are considered upward and downward "reclassifications".

The IDI is defined as

$$\hat{\text{IDI}} = (\bar{\hat{p}}_{\text{new, events}} - \bar{\hat{p}}_{\text{new, nonevents}}) - (\bar{\hat{p}}_{\text{old, events}} - \bar{\hat{p}}_{\text{old, nonevents}}),$$

where $\bar{\hat{p}}_{\text{new, events}}$ is the mean of predicted probabilities of the event, based on the new model for individuals who experience the event and $\bar{\hat{p}}_{\text{new, nonevents}}$ is the mean of predicted probabilities of the event, based on the new model for individuals who do not experience an event. The probabilities $\bar{\hat{p}}_{\text{old, events}}$ and $\bar{\hat{p}}_{\text{old, nonevents}}$ are calculated respectively using the old model.

The NRI and IDI have recently been criticized (Hilden, 2014; Pepe et al., 2014; Vickers and Pepe, 2014) and so-called decision-analytic measures have been recommended to be used instead, because they are considered to be clinically more meaningful

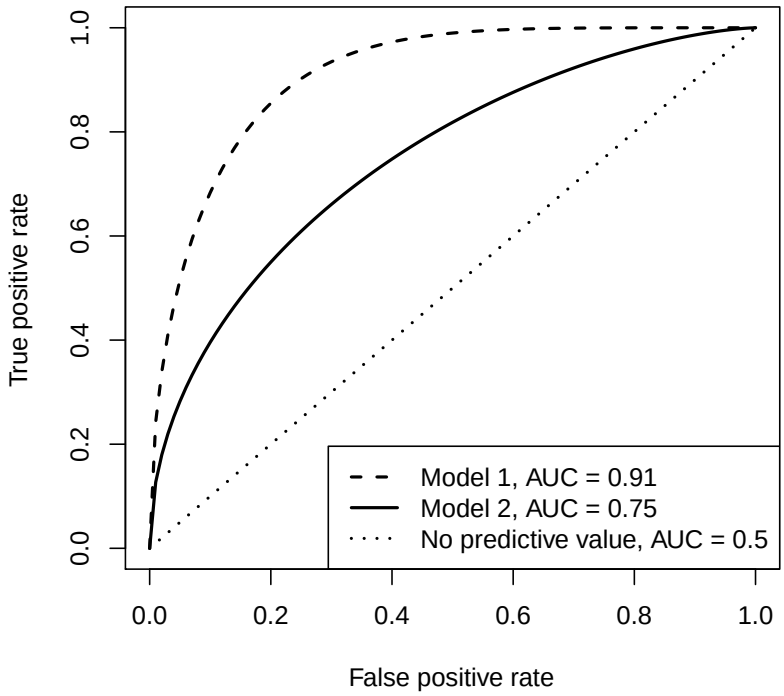

Figure 2.1: An example of the ROC curves of two models. Model 1 has a larger area under the curve (AUC) than Model 2, which means that Model 1 has a better discriminative ability.

(Van Calster et al., 2013; Kerr et al., 2014). In decision-analytic measures, the relative consequences of false positives and negatives are taken into account by a threshold T for predicted risk, which is used to categorize individuals as positive or negative. The threshold T is defined so that the odds $T/(1-T)$ equals to the ratio of the harm of a FP decision to the benefit of a TP decision. Once the threshold T is defined, the model performance can be measured using so called net benefit (NB) (Vickers and Elkin, 2006), which is given by

$$\mathrm{NB} = \frac{\mathrm{TP}}{n} - w\frac{\mathrm{FP}}{n},$$

where n is the total number of individuals and $w = T/(1-T)$. The difference in NB between two models can be interpreted as the difference in the proportion of true positives at the same level of false positives.

Model calibration is often measured using the Hosmer–Lemeshow goodness-of-fit test (Hosmer and Lemeshow, 1980; Hosmer et al., 2013). The predicted risks are categorized to, for example, deciles with respect to their values, and then the observed and

expected numbers of events are compared in the subgroups. The Hosmer–Lemeshow statistic is defined as

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - n_g\bar{\pi}_g)^2}{n_g\bar{\pi}_g(1-\bar{\pi}_g)},$$

where G is the number of subgroups, O_g is the number of events in the gth group, n_g is the number of individuals in the gth group and $\bar{\pi}_g$ is the average predicted risk for individuals in group g. The statistic χ^2_{HL} follows asymptotically a χ^2 distribution with $g-2$ degrees of freedom. Calibration can also be evaluated using a calibration plot and the measures related to it: calibration slope and calibration-in-the-large (Steyerberg et al., 2010).

In addition to model discrimination and calibration, there are also measures indicating the overall model performance, such as the Brier score (BS) (Brier, 1950). The BS quantifies how close the predictions are to actual outcomes. For binary outcomes the BS is written

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - O_i)^2,$$

where $\hat{\pi}_i$ is the predicted risk and O_i is the outcome of the binary response for individual i. The metrics discussed above have several extensions and modifications for different types of data and models (Li and Fine, 2008; Chambless et al., 2011; Demler et al., 2015).

Article III demonstrates that different measures of model performance can be used to plan how many longitudinal covariate measurements are needed. In particular, this is based on application of these measures to compare models with different amounts of longitudinal information. The contribution of Article III is summarized in Section 4.3.

A different aspect in assessing the goodness of a survival model is checking the validity of the model assumptions. For example, in parametric models, the choice of the distribution for survival times can be checked (Hollander and Proschan, 1979; Collett, 2003). For assessing the validity of the assumptions in proportional hazards models, several residual-based procedures have been developed, including the Schoenfeld residuals (Schoenfeld, 1982), the Cox-Snell residuals (Cox and Snell, 1968), and martingale residuals (Barlow and Prentice, 1988). Collett (2003) provides a discussion on the use of different residuals in examining the adequacy of the linear component $\boldsymbol{x}^T\boldsymbol{\beta}$ of a survival model and the validity of the assumption of proportional hazards.

Chapter 3

Optimal study design

The selected study design affects on how well the parameters of interest can be estimated. Different ways of data collection may also require different amounts of resources. Thus, it is often of interest to consider how to design the study in order to achieve the desired precision with costs as low as possible or to maximize the precision of results with a fixed budget. This chapter presents some concepts and methods for study design optimization.

3.1 Optimal experimental design methods in observational studies

This thesis deals with the cost-efficient planning of observational studies. In Articles I and II, the principles of optimal experimental design are used. This methodology will be reviewed in Sections 3.2–3.4. The central difference in constructing optimal designs in experimental and observational studies is that in experiments the values of the covariates (or design points) for each subject can be determined, whereas in observational studies this is not possible.

When the covariate values cannot be determined, we have to select the subjects whose covariate values can be expected to equal the desired values. Alternatively, the expected covariate values of the candidate subjects can be thought to be the set of possible design points, from which the optimal design will be constructed. In practice, the calculation of the expected values requires some prior knowledge or assumptions about the processes that generate the covariates. For this reason, optimal design methods may be used in longitudinal studies, for instance, where information on the same variables is collected repeatedly and the study can be designed in different phases. Applications of optimal design methods in subject selection problems are considered, e.g., by Karvanen et al. (2009) and Buzoianu and Kadane (2009). Such a selection problem is also studied in Articles I and II, whose contribution is summarized in Sections 4.1 and 4.2.

There are also other ways to apply optimal design methods in observational studies than the optimal subject selection described above. These may be the determination of

the optimal number of repeated measurements or the optimal time spacing of measurements. Tekle et al. (2011) study these questions in the case of continuous longitudinal response under budget constraints. Optimal time spacing has also been investigated for categorical processes in various settings (Hwang and Brookmeyer, 2003; Quintana and Müller, 2004; Mehtälä et al., 2015).

3.2 Optimal design theory

The optimality of a study design can be approached from many different viewpoints. The aim of optimization may be, for example, maximization of precision for parameter estimates or predictions, minimization of costs or time required to conduct the study or a multi-objective function combining different goals. In epidemiology, designs such as case-control design, nested case-control design, and case-cohort design (Wacholder, 1991; Kulathinal et al., 2007; Sun et al., 2010; Salim et al., 2014) have been used to improve the cost-efficiency of studies. Optimal survey sampling techniques have been developed in survey statistics (Chambers and Clark, 2012), and the so-called alphabetic optimality theory (Atkinson et al., 2007), in which different optimality criteria are identified with letters, has been used in experimental research. The treatment of the topic is here restricted to the alphabetic optimality, which is presented in this section, and to the Bayesian experimental design framework summarized in Section 3.3.

Let us consider a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. The amount of information on parameters $\boldsymbol{\theta}$ can be expressed as a Fisher information matrix $I(\boldsymbol{\theta})$, from which we are able to derive a confidence ellipsoid for $\boldsymbol{\theta}$ under a normal distribution assumption for the estimates. Some of the optimality criteria concern different properties of confidence ellipsoids. A D-optimal design, which minimizes the volume of the ellipsoid, is obtained by maximizing $\det(I(\boldsymbol{\theta}))$ or, equivalently, by minimizing the generalized variance $\det(I(\boldsymbol{\theta})^{-1})$. A-optimality is defined as minimizing the $\text{trace}(I(\boldsymbol{\theta})^{-1})$, which results in minimizing the total variance of the estimates, equivalent to minimizing the average variance. An E-optimal design minimizes the variance of the least well estimated linear combination $\boldsymbol{a}^T\boldsymbol{\theta}$, where $\boldsymbol{a}$ is a vector of coefficients and $\boldsymbol{a}^T\boldsymbol{a} = 1$.

The three criteria introduced above can also be defined considering the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $I(\boldsymbol{\theta})$. In fact, the definitions make use of the eigenvalues of $I(\boldsymbol{\theta})^{-1}$, which are $1/\lambda_1, \ldots, 1/\lambda_p$, and are proportional to the squared lengths of the axes of the confidence ellipsoid. Now, the D-, A- and E-optimal designs are obtained by minimizing $\prod_{i=1}^{p} 1/\lambda_i$, $\sum_{i=1}^{p} 1/\lambda_i$ and $\max_i(1/\lambda_i)$, respectively (Atkinson et al., 2007).

If we are only interested in a subset of s parameters, a variation of the D-optimality called the D_s-optimality criterion can be applied. Assume that the parameters have been ordered so that the parameters of interest are the first s elements in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s, \ldots, \theta_p)^T$. The D_s-criterion is defined as minimizing the determinant of the $s \times s$ upper left submatrix of $I(\boldsymbol{\theta})^{-1}$.

D-optimal design has the advantage that it does not depend on the scales of the variables, even though $I(\boldsymbol{\theta})$ does. Thus, if A-optimality is used, for instance, it may be appropriate to scale the variables to have the same variance.

A complex issue is encountered with non-linear models, when $I(\boldsymbol{\theta})$, and hence the optimal design, depends on the model parameters, which are usually unknown. Replacing the unknown parameters with prior point estimates leads to locally optimal designs. Taking into account the uncertainty about the parameters by using prior distributions instead of point estimates is discussed in the next section.

In some cases, it is possible to improve the study design by considering sequential (or multi-phase) designs. This means that the entire data collection is not designed at once but in different phases. This allows us to utilize the information already collected in the study in order to design the next phase possibly even more efficiently. Two-phase designs are important special cases of sequential designs and have been widely examined to improve the cost-efficiency of studies (Breslow and Cain, 1988; Reilly, 1996; McIsaac and Cook, 2015). Sequential designs may be particularly useful for non-linear models, due to the dependency between the optimal design and the unknown model parameters.

3.3 Bayesian optimal design

The Bayesian approach may be natural in design optimization, as we usually have some prior information which has motivated conducting a new study. On the other hand, deriving optimal designs for non-linear models requires prior information on model parameters, and the information can often be incorporated flexibly using prior distributions. Bayesian experimental design (Chaloner and Verdinelli, 1995) is a framework in which the aim is to find a design ξ from a design space Ξ, which maximizes the expected utility $U(\xi)$ of the experiment. The utility should describe the purpose of the experiment and can be defined to be the precision of the parameter estimates or predictions, for instance, analogously with the previous section.

Data $\boldsymbol{y}$ from a sample space $\mathcal{Y}$ are collected according to design ξ. The data are assumed to follow a model $p(\boldsymbol{y}|\boldsymbol{\theta})$, where parameters $\boldsymbol{\theta}$ belong to the parameter space Θ. The observed utility is measured by a utility function $u(\boldsymbol{\theta}, \xi, \boldsymbol{y})$. The Bayesian solution to the design problem is obtained by finding the design ξ^*, so that

$$U(\xi^*) = \max_{\xi \in \Xi} \int_{\mathcal{Y}} \int_{\Theta} u(\boldsymbol{\theta}, \xi, \boldsymbol{y}) p(\boldsymbol{\theta}, \boldsymbol{y}|\xi) d\boldsymbol{\theta} d\boldsymbol{y}. \tag{3.1}$$

The integrals in (3.1) average over what is unknown: data $\boldsymbol{y}$ have not yet been collected and for the parameters $\boldsymbol{\theta}$ only a prior distribution is assumed.

When the goal is to obtain precise estimates for the parameters $\boldsymbol{\theta}$, it is common to use the expected Shannon information of the posterior distribution as the expected utility to be maximized:

$$U(\xi) = \int_{\mathcal{Y}} \int_{\Theta} \log\left[p(\boldsymbol{\theta}|\boldsymbol{y}, \xi)\right] p(\boldsymbol{\theta}, \boldsymbol{y}|\xi) d\boldsymbol{\theta} d\boldsymbol{y}.$$

This is equivalent to maximizing the expected Kullback–Leibler distance between the posterior and the prior distributions, and it leads to Bayesian D-optimality under a

normal linear model. In practice, the integrals in (3.1) are often replaced with approximations (Atkinson et al., 1995; Ryan, 2003). Article I applies the D- and D_s-optimality criteria in frequentist framework, and in Article II we use Bayesian version of D_s-optimality.

3.4 Optimal design for survival models

The application of optimal design methods in survival models has received relatively little attention (McGree and Eccleston, 2010). Due to the non-linearity of survival models, the Bayesian approach has often been employed (Erkanli and Soyer, 2000; Zhang and Meeker, 2006). McGree and Eccleston (2010) investigate optimal covariate values for precise parameter estimation in frailty models and present a compound criterion, which aims at maximizing simultaneously the precision of parameter estimates and the number of failures in an experiment.

Rivas-López et al. (2014) study optimal designs for the precise estimation of model parameters in survival experiments using AFT models. They show that in a case of one covariate with a linear effect, the D-optimal design is based on setting two covariate values. When the interval of possible covariate values is $[a, b]$ and the regression parameter $\beta > 0$ (the greater the covariate value, the longer the expected survival), the design points are the minimum a and the so-called critical point $c > a$. If $c \notin [a, b]$, then the design will consist of the extreme points a and b. Respectively, when $\beta < 0$, the design points are b and $c' < b$. The authors suppose that the explanation for such critical points produced by the optimality criterion is that using them instead of the extreme point where the expected survival time is the longest results in more events that give more information than censored observations.

The majority of research on optimal design for survival models deals with parametric models. Although fitting semiparametric Cox models using partial likelihood is rather straightforward, design optimization based on partial likelihood information is a complicated task. Recently, López-Fidalgo and Rivas-López (2014) and Konstantinou et al. (2015) have presented methods for finding optimal covariate values for the Cox model. The designs based on partial likelihood seem to be similar to designs based on the corresponding full likelihood (Konstantinou et al., 2015).

Chapter 4

Research contribution

This section summarizes the contribution of Articles I–IV to the research questions outlined in Chapter 1.

4.1 Optimal subcohort selection

Article I considers the selection of individuals for a re-measurement of a single time-varying covariate when only a subset of the cohort can be selected due to budget limitations. It is shown that in order to obtain a precise estimate of the covariate effect in a survival model (2.5), the oldest individuals with extreme covariate values should be selected. The proposed selection methods are based on functions of the Fisher information matrix presented in Section 3.2, and the results indicate that these methods lead to more precise estimates than simple random sampling. Two different approaches of Section 2.4 are used in the handling of missing data: multiple imputation and a likelihood-based approach with numerical integration. Numerical integration is seen to perform well in a simulation study, but due to its sensitivity to model assumptions, multiple imputation is recommended for analysis of real data.

4.2 Extensions of the optimal subcohort selection

The selection procedure introduced in Article I is generalized in Article II to allow for several covariates and measurement points. The Bayesian approach introduced in Section 3.3 is applied here and is found to be suitable for this kind of problem of sequential study design, both from the theoretical and practical points of view. The results are consistent with Article I: the optimal subcohort consists of old individuals with extreme covariate values and the proposed selection method clearly outperforms simple random sampling, when the precision of the regression parameters is compared. Bayesian data augmentation appears to be a more flexible method for the handling of missing data than the methods applied in Article I.

4.3 Frequency of longitudinal measurements

In Article III, we study how the number of longitudinal covariate measurements can be chosen cost-efficiently by evaluating the usefulness of the measurements for risk prediction. Simulations as well as data from a previous study were used to illustrate the importance of longitudinal measurements. We propose applying measures of model discrimination, presented in Section 2.5, to compare models using different amounts of longitudinal information. By performing analyses considering the usefulness of longitudinal covariate measurements, we are able to conclude to what extent we could decrease the number of the measurements without significantly losing the precision of the predictions. These kind of comparisons seem applicable in both ongoing and completely new follow-up studies. In a simulation study, we show how higher variability and a higher hazard ratio of a time-varying covariate increase the importance of re-measurements.

4.4 Modeling of cardiovascular disease mortality

The use of longitudinal covariate measurements in modeling the risk of cardiovascular disease (CVD) mortality in a long-term follow-up study is investigated in Article IV. The research is based on the Finnish cohorts of the international Seven Countries Study. New variables from the longitudinal measurements are derived as discussed in Section 2.3, and their importance in statistical modeling is analysed. Changes in the covariate values are modeled as the difference of the latest two measurements and the cumulative average as a mean of the most recent and all previous measurements. The use of these new time-varying covariates is compared to the traditional use of time-varying covariates, in which only the most recent measurement is assumed to affect the risk.

Individual level changes and cumulative values are strong predictors for cardiovascular disease mortality. In particular, the long-term cumulative value of systolic blood pressure is found to be a better predictor than the recent level alone. The change in the body mass index predicted the risk of CVD mortality although the body mass index itself did not. The article confirms the value of longitudinal risk factor information in risk prediction. Moreover, we conclude that using a simplistic method in handling longitudinal risk factor measurements in a prediction model may prevent researchers from understanding the true importance of the risk factors.

Chapter 5

Summary

This thesis dealt with efficient designs and modeling approaches for follow-up studies where time-varying covariates are present. We studied the optimal selection of a subcohort for re-measurements of the covariates in order to estimate the covariate effects on survival as precisely as possible (Article I and II). Another perspective to cost-efficiency of follow-up studies was to consider the determination of the reasonable number of longitudinal measurements for risk prediction (Article III). Different ways to utilize longitudinal covariate measurements in modeling CVD mortality were investigated in Article IV.

The problem of the optimal selection of a subcohort was solved using optimality criteria developed for the design of experiments, and this approach was shown to outperform simple random sampling. The optimal subcohorts consisted of individuals with extreme covariate values and high age. For choosing the number of longitudinal measurements, we proposed utilizing data from previous studies and/or simulations and applying measures of model discrimination to compare models using different amounts of longitudinal information. Finally, we showed that individual-level changes and cumulative averages of classical risk factors are good predictors of CVD mortality.

There has been relatively little research on the application of methods developed for the optimal design of experiments to the design of observational studies. However, our results encourage to further investigate this approach in study design. Although the applications of this thesis were in epidemiology, similar approaches could be used in other disciplines as well.

The results demonstrated that the cost-efficiency of follow-up designs can be improved by careful planning compared to simple solutions. This work may help researchers to conduct follow-up studies with time-varying covariates more efficiently. We also showed that taking full advantage of the collected follow-up data with longitudinal measurements may require deriving new covariates and comparing their use with simpler methods.

In addition to simulations, the methods proposed for study design were also studied with real data from the East–West study. This gives more reliability for the applicability of the methods, as the processes which have generated the data are truly unknown.

Nevertheless, our assumptions may not be valid with other data sets, and so caution is required when generalizing our conclusions to other studies. It is also worth noting as a limitation of our methods that the study is designed with respect to a specific model, and it may thus give reduced power for fitting models with other variables. Thus the proposed methods for designing a study are at their best when the use of the data can be clearly defined before the data collection.

This dissertation considered only some aspects of cost-efficiency in follow-up studies and the utilization of collected data in modeling. More work is needed to discover the possibilities of design optimization in follow-up studies with time-varying covariates. A topic for future research could be combining the methods and concepts of Articles I–IV in different applications. Another interesting topic would be to explore the optimal combination of the number of individuals in the cohort and the number of longitudinal measurement for each individual. It was assumed in this work that the size of the original cohort is fixed in advance and that the sizes of the subcohorts are the same in each re-measurement. Relaxing these assumptions might allow the designs to be further improved.

Bibliography

Aalen, O., Borgan, Ø., and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York.

Atkinson, A., Demetrio, C., and Zocchi, S. (1995). Optimum dose levels when males and females differ in response. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(2):213–226.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford.

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, 75(1):65–74.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99.

Breslow, N. and Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2):221–228.

Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73.

Buzoianu, M. and Kadane, J. B. (2009). Optimal Bayesian design for patient selection in a clinical study. *Biometrics*, 65(3):953–961.

Bycott, P. and Taylor, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine*, 17(18):2061–2077.

Carneiro, I. and Howard, N. (2011). *Introduction to Epidemiology*. Open University Press, Maidenhead, UK.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10(3):273–304.

Chambers, R. and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford.

Chambless, L. E., Cummiskey, C. P., and Cui, G. (2011). Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in Medicine*, 30(1):22–38.

Chiou, S. H., Kang, S., Kim, J., and Yan, J. (2014). Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. *Lifetime Data Analysis*, 20(4):599–618.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. CRC Press, Boca Raton, 2nd edition.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 30(2):248–275.

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462.

Dekker, F. W., de Mutsert, R., van Dijk, P. C., Zoccali, C., and Jager, K. J. (2008). Survival analysis: time-dependent effects and time-varying risk factors. *Kidney International*, 74(8):994–997.

Demler, O. V., Paynter, N. P., and Cook, N. R. (2015). Tests of calibration and goodness-of-fit in the survival setting. *Statistics in Medicine*, 34(10):1659–1680.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Erkanli, A. and Soyer, R. (2000). Simulation-based designs for accelerated life tests. *Journal of Statistical Planning and Inference*, 90(2):335–348.

Euser, A. M., Zoccali, C., Jager, K. J., and Dekker, F. W. (2009). Cohort studies: prospective versus retrospective. *Nephron Clinical Practice*, 113(3):c214–c217.

Farchi, G., Capocaccia, R., Verdecchia, A., and Menotti, A. (1981). Risk factors changes and coronary heart disease in an observational study. *International Journal of Epidemiology*, 10(1):31–40.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Gautschi, W. (2012). *Numerical Analysis*. Birkhäuser, New York, 2nd edition.

Gould, L. A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*, 34(14):2181–2195.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hanson, T. E., Branscum, A. J., and Johnson, W. O. (2011). Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches. *Lifetime Data Analysis*, 17(1):3–28.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hilden, J. (2014). Commentary: On NRI, IDI, and "good-looking" statistics with nothing underneath. *Epidemiology*, 25(2):265–267.

Hollander, M. and Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics*, 35(2):393–401.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, New Jersey, 3rd edition.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037–1043.

Hwang, W.-T. and Brookmeyer, R. (2003). Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis*, 9(3):261–274.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, 2nd edition.

Karvanen, J., Kulathinal, S., and Gasbarra, D. (2009). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, 53(5):1782–1793.

Karvonen, M. J., Blomqvist, G., Kallio, V., Orma, E., Punsar, S., Rautaharju, P., Takkunen, J., and Keys, A. (1966). Men in rural East and West Finland. *Acta Medica Scandinavica*, 180(s460):169–190.

Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., and Pepe, M. S. (2014). Net reclassification indices for evaluating risk prediction instruments: A critical review. *Epidemiology*, 25(1):114–121.

Keys, A. (1970). Coronary heart disease in seven countries. *Circulation*, 41(1):186–195.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition.

Kom, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, 145(1):72–80.

Konstantinou, M., Biedermann, S., and Kimber, A. C. (2015). Optimal designs for full and partial likelihood information – with application to survival models. *Journal of Statistical Planning and Inference*, 165:27–37.

Krickeberg, K., Pham, V. T., and Pham, T. M. H. (2012). *Epidemiology: Key to Prevention*. Springer, New York.

Kulathinal, S., Karvanen, J., Saarela, O., Kuulasmaa, K., and for the MORGAM Project (2007). Case-cohort design in practice – experiences from the MORGAM Project. *Epidemiological Perspectives & Innovations*, 4(1):15.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624–632.

Li, J. and Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3):566–576.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York, 2nd edition.

López-Fidalgo, J. and Rivas-López, M. (2014). Optimal experimental designs for partial likelihood information. *Computational Statistics & Data Analysis*, 71:859–867.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press, Boca Raton.

McGree, J. and Eccleston, J. (2010). Investigating design for survival models. *Metrika*, 72(3):295–311.

McIsaac, M. A. and Cook, R. J. (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in Medicine*, 34(21):2899–2912.

Mehtälä, J., Auranen, K., and Kulathinal, S. (2011). Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Statistical Methods in Medical Research*, doi: 10.1177/0962280211430663.

Mehtälä, J., Auranen, K., and Kulathinal, S. (2015). Optimal observation times for multistate Markov models – applications to pneumococcal colonization studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):451–468.

Menotti, A. and Puddu, P. E. (2013). Coronary heart disease differences across Europe: a contribution from the Seven Countries Study. *Journal of Cardiovascular Medicine*, 14(11):767–772.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Muntner, P., Shimbo, D., Tonelli, M., Reynolds, K., Arnett, D. K., and Oparil, S. (2011). The relationship between visit-to-visit variability in systolic blood pressure and all-cause mortality in the general population: Findings from NHANES III, 1988 to 1994. *Hypertension*, 57(2):160–166.

Pencina, M. J., D'Agostino, R. B., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, 30(1):11–21.

Pencina, M. J., D'Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172.

Pepe, M. S., Fan, J., Feng, Z., Gerds, T., and Hilden, J. (2014). The net reclassification index (NRI): A misleading measure of prediction improvement even with independent test data sets. *Statistics in Biosciences*, doi: 10.1007/s12561-014-9118-0.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational Statistics & Data Analysis*, 53(4):1142–1154.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley, New York.

Quintana, F. A. and Müller, P. (2004). Optimal sampling for repeated binary measurements. *The Canadian Journal of Statistics*, 32(1):73–84.

Reilly, M. (1996). Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology*, 143(1):92–100.

Rivas-López, M. J., López-Fidalgo, J., and Campo, R. d. (2014). Optimal experimental designs for accelerated failure time with Type I and random censoring. *Biometrical Journal*, 56(5):819–837.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press, Boca Raton.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12):1366–1380.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.

Rothwell, P. M., Howard, S. C., Dolan, E., O'Brien, E., Dobson, J. E., Dahlöf, B., Sever, P. S., and Poulter, N. R. (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet*, 375(9718):895–905.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in probability and mathematical statistics. John Wiley & Sons, Inc., New York.

Ryan, K. J. (2003). Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, 12(3):585–603.

Salim, A., Ma, X., Fall, K., Andrén, O., and Reilly, M. (2014). Analysis of incidence and prognosis from 'extreme' case-control designs. *Statistics in Medicine*, 33(30):5388–5398.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Sesso, H. D., Stampfer, M. J., Rosner, B., Gaziano, J. M., and Hennekens, C. H. (2000). Two-year changes in blood pressure and subsequent risk of cardiovascular disease in men. *Circulation*, 102(3):307–312.

Shen, H. and Cook, R. J. (2013). Regression with incomplete covariates and left-truncated time-to-event data. *Statistics in Medicine*, 32(6):1004–1015.

Smithson, M. and Merkle, E. C. (2014). *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. CRC Press, Boca Raton.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138.

Sun, W., Joffe, M. M., Chen, J., and Brunelli, S. M. (2010). Design and analysis of multiple events case-control studies. *Biometrics*, 66(4):1220–1229.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tekle, F. B., Tan, F. E., and Berger, M. P. (2011). Too many cohorts and repeated measurements are a waste of resources. *Journal of Clinical Epidemiology*, 64(12):1383–1390.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Thiébaut, A. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–3820.

Tsiatis, A., De Gruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Van Calster, B., Vickers, A. J., Pencina, M. J., Baker, S. G., Timmerman, D., and Steyerberg, E. W. (2013). Evaluation of markers and risk prediction models: Overview of relationships between NRI and decision-analytic measures. *Medical Decision Making*, 33(4):490–501.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Vickers, A. J. and Pepe, M. (2014). Does the net reclassification improvement help us evaluate models and markers? *Annals of Internal Medicine*, 160(2):136–137.

Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, 2(2):155–158.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Wilson, P. W., Hoeg, J. M., D'Agostino, R. B., Silbershatz, H., Belanger, A. M., Poehlmann, H., O'Leary, D., and Wolf, P. A. (1997). Cumulative effects of high cholesterol levels, high blood pressure, and cigarette smoking on carotid stenosis. *New England Journal of Medicine*, 337(8):516–522.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Zhang, Y. and Meeker, W. Q. (2006). Bayesian methods for planning accelerated life tests. *Technometrics*, 48(1):49–60.

INCLUDED ARTICLES

I

Optimal selection of individuals for repeated covariate measurements in follow-up studies

Jaakko Reinikainen, Juha Karvanen and Hanna Tolonen

Statistical Methods in Medical Research, 2014,
DOI: 10.1177/0962280214523952.

II

Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies

Jaakko Reinikainen and Juha Karvanen

Submitted, 2015

III

How many longitudinal covariate measurements are needed for risk prediction?

Jaakko Reinikainen, Juha Karvanen and Hanna Tolonen

Journal of Clinical Epidemiology, 2015,
DOI: 10.1016/j.jclinepi.2015.06.022.

IV

Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland

Jaakko Reinikainen, Tiina Laatikainen, Juha Karvanen and Hanna Tolonen

International Journal of Epidemiology, 2015, 44(1): 108-116.

120. MYLLYMÄKI, MARI, Statistical models and inference for spatial point patterns with intensity-dependent marks. (115 pp.) 2009
121. AVIKAINEN, RAINER, On generalized bounded variation and approximation of SDEs. (18 pp.) 2009
122. ZÜRCHER, THOMAS, Regularity of Sobolev–Lorentz mappings on null sets. (13 pp.) 2009
123. TOIVOLA, ANNI, On fractional smoothness and approximations of stochastic integrals. (19 pp.) 2009
124. VIHOLA, MATTI, On the convergence of unconstrained adaptive Markov chain Monte Carlo algorithms. (29 pp.) 2010
125. ZHOU, YUAN, Hajłasz–Sobolev extension and imbedding. (13 pp.) 2010
126. VILPPOLAINEN, MARKKU, Recursive set constructions and iterated function systems: separation conditions and dimension. (18 pp.) 2010
127. JULIN, VESA, Existence, uniqueness and qualitative properties of absolute minimizers. (13 pp.) 2010
128. LAMMI, PÄIVI, Homeomorphic equivalence of Gromov and internal boundaries. (21 pp.) 2011
129. ZAPADINSKAYA, ALEKSANDRA, Generalized dimension distortion under Sobolev mappings. (18 pp.) 2011
130. KEISALA, JUKKA, Existence and uniqueness of $p(x)$-harmonic functions for bounded and unbounded $p(x)$. (56 pp.) 2011
131. SEPPÄLÄ, HEIKKI, Interpolation spaces with parameter functions and L_2-approximations of stochastic integrals. (18 pp.) 2011
132. TUHOLA-KUJANPÄÄ, ANNA, On superharmonic functions and applications to Riccati type equations. (17 pp.) 2012
133. JIANG, RENJIN, Optimal regularity of solutions to Poisson equations on metric measure spaces and an application. (13 pp.) 2012
134. TÖRMÄKANGAS, TIMO, Simulation study on the properties of quantitative trait model estimation in twin study design of normally distributed and discrete event-time phenotype variables. (417 pp.) 2012
135. ZHANG, GUO, Liouville theorems for stationary flows of generalized Newtonian fluids. (14 pp.) 2012
136. RAJALA, TUOMAS, Use of secondary structures in the analysis of spatial point patterns. (27 pp.) 2012
137. LAUKKARINEN, EIJA, On Malliavin calculus and approximation of stochastic integrals for Lévy processes. (21 pp.) 2012
138. GUO, CHANGYU, Generalized quasidisks and the associated John domains. (17 pp.) 2013
139. ÄKKINEN, TUOMO, Mappings of finite distortion: Radial limits and boundary behavior. (14 pp.) 2014
140. ILMAVIRTA, JOONAS, On the broken ray transform. (37 pp.) 2014
141. MIETTINEN, JARI, On statistical properties of blind source separation methods based on joint diagonalization. (37 pp.) 2014
142. TENGVALL, VILLE, Mappings of finite distortion: Mappings in the Sobolev space $W^{1,n-1}$ with integrable inner distortion. (22 pp.) 2014
143. BENEDICT, SITA, Hardy-Orlicz spaces of quasiconformal mappings and conformal densities. (16 pp.) 2014
144. OJALA, TUOMO, Thin and fat sets: Geometry of doubling measures in metric spaces. (19 pp.) 2014
145. KARAK, NIJJWAL, Applications of chaining, Poincaré and pointwise decay of measures. (14 pp.) 2014
146. JYLHÄ, HEIKKI, On generalizations of Evans and Gangbo's approximation method and L^∞ transport. (20 pp.) 2014
147. KAURANEN, AAPO, Space-filling, energy and moduli of continuity. (16 pp.) 2015
148. YLINEN, JUHA, Decoupling on the Wiener space and variational estimates for BSDEs. (45 pp.) 2015
149. KIRSILÄ, VILLE, Mappings of finite distortion on generalized manifolds. (14 pp.) 2015
150. XIANG, CHANG-LIN, Asymptotic behaviors of solutions to quasilinear elliptic equations with Hardy potential. (20 pp.) 2015
151. ROSSI, EINO, Local structure of fractal sets: tangents and dimension. (16 pp.) 2015
152. HELSKE, JOUNI, Prediction and interpolation of time series by state space models. (28 pp.) 2015

ISBN 978-951-39-6315-6
ISSN 1457-8905