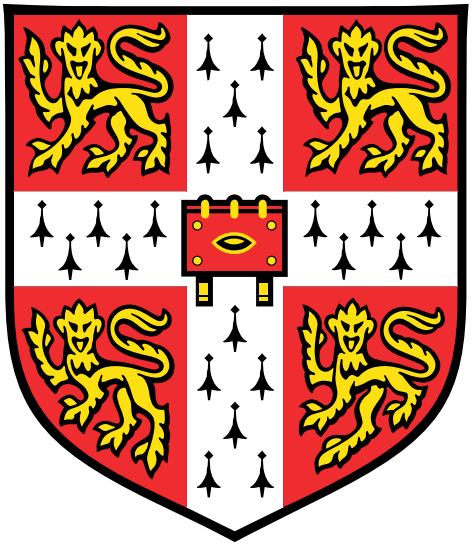
University of Cambridge



Beyond Parameter Estimation: Analysis of the Case-Cohort Design in Cox Models

Susan E. Connolly
St John's College

May 2019

This dissertation is submitted for the degree of *Doctor of Philosophy*

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

Beyond Parameter Estimation: Analysis of the Case Cohort Design in Cox Models
Susan E. Connolly

Cohort studies allow for powerful analysis, but an exposure may be too expensive to measure in the whole cohort. The case-cohort design measures covariates in a random sample (subcohort) of the full cohort, as well as in all cases that emerge, regardless of their initial presence in the subcohort. It is an increasingly popular method, particularly for medical and biological research, due to its efficiency and flexibility. However, the case-cohort design poses a number of challenges for estimation and post-estimation procedures. Cases are over-represented in the dataset, and hence estimation of coefficients in this design requires weighting of observations. This results in a pseudopartial likelihood, and standard post-estimation methods may not be readily transferable to the case-cohort design.

This thesis presents theory and simulation studies for application of estimation and post-estimation methods in the case-cohort design. In the majority of extant literature considering methods for the case-cohort design, simulation studies generally consider full cohort sizes, sampling fractions, and case percentages that are dissimilar to those seen in practice. In this thesis the design of the simulation studies aims to provide circumstances which are similar to those encountered when using case-cohort designs in practice. Further, these methods are applied to the InterAct dataset, and practical advice and sample code for STATA is presented.

Estimation of Coefficients & Cumulative Baseline Hazard: For estimation of coefficients, Prentice weighting and Barlow weighting are the most commonly used (Sharp et al, 2014). Inverse Probability Weighting (IPW), in this context, refers to methods where the entire case-cohort sample at risk is used in the analysis, as opposed to Prentice and Barlow weighting systems, where cases outside the subcohort sample are only included in risk sets just prior to their time of failure. This thesis assesses bias and precision of Prentice, Barlow and IPW weighting methods in

the case-cohort design. Simulation studies show IPW, Prentice and Barlow weighting to have similar low bias. Where case percentage is high, IPW weighting shows an increase in precision over Prentice and Barlow, though this improvement is small.

Checks of Model Assumptions: Appropriateness of covariate functional form in the standard Cox model can be assessed graphically by smoothed martingale residuals against various other values, such as time and covariates of interest (Therneau et al, 1990). The over-representation of cases in the case-cohort data, as compared to the full cohort, distorts the properties of such residuals. Methods related to IPW that adapt such plots to the case-cohort design are presented. Detection of non-proportional hazards by use of Schoenfeld residuals, scaled Schoenfeld residuals, and inclusion of time-varying covariates in the model are assessed and compared by simulation studies, finding that where risk set sizes are not overly variable, all three methods are appropriate for use in the case-cohort design, with similar power. Where case-cohort risk set sizes are more variable, methods based on Schoenfeld residuals and scaled Schoenfeld residuals show high Type 1 error rate.

Model Comparison & Variable Selection: The methods of Lumley & Scott (2013, 2015) for modification of the Likelihood Ratio test (dLR), AIC ($dAIC$) and BIC ($dBIC$) in complex survey sampling are applied to case-cohort data and assessed in simulation studies. In the absence of sparse data, dLR is found to have similar power to robust Wald tests, with Type 1 error rate approximately 5%. In the presence of sparse data, the dLR is superior to robust Wald tests. In the absence of sparse data $dBIC$ shows little difference from the naïeve use of the pseudo-log-likelihood in the standard BIC formula ($pBIC$). In the presence of sparse data $dBIC$ shows reduced power to select the true model, and $pBIC$ is superior. $dAIC$ shows improvement in power to select the true model over naïeve methods. Where subcohort size and number of cases is not overly small, loss of power from the full cohort for $dAIC$, $dBIC$ and $pBIC$ is not substantial.

Acknowledgements

Above all I must acknowledge the incomparable support and advice given by my supervisor, Dr. Ian White. Ian is a paragon of the best of academia; collegiality, search for knowledge, and scientific rigour. It has been a privilege to be supervised by such an excellent mentor.

My advisors, Dr. Paul Newcombe, Dr. Stephen Sharp, and Dr. Angela Wood consistently gave excellent guidance and advice, and I thank them sincerely for their efforts and patience. I thank the students and staff of the MRC Biostatistics Unit for fostering a stimulating and supportive research environment, and St. John's College Cambridge for being a welcoming home for the last 3 years.

The EPIC-InterAct study received funding from the European Union (Integrated Project LSHM-CT-2006-037197 in the Framework Programme 6 of the European Community). I thank all participants and staff for their contribution to this study. I thank the EPIC-InterAct PI, management team and wider consortium for their permission to use the data, and Nicola Kerrison (MRC Epidemiology Unit, University of Cambridge) for preparing the dataset which I used in Chapters 2 and 7. I acknowledge personal financial support from the UK Medical Research Council and St John's College, Cambridge.

This thesis would not have been possible without the support of my family, and that of my friends, especially the members of LWSC, CFSODC, and JSPNO in Ireland, and the wonderful new friends I found in Cambridge. This thesis is dedicated to those friends and family, who have always seen the best in me.

Contents

List of Figures

List of Tables

Notation

The following notation is used throughout this dissertation. This is intended as a reference and, along with all additional notation, shall be properly introduced within the main body of the document.

$t_{(j)}$	Failure time for the jth subject in an ordered list of failure times
$t_{(0)i}$	Entry time of the ith subject
t_i	Recorded survival time of the ith subject
$R_{(j)}$	Set of observations at risk at time $t_{(j)}$ in the full cohort, not including the failure at time $t_{(j)}$
$R^C_{(j)}$	The subset of $R_{(j)}$ consisting only of cases
$R^{NC}_{(j)}$	The subset of $R_{(j)}$ consisting only of non-cases
$R^*_{(j)}$	Set of observations at risk at time $t_{(j)}$ in the case cohort, not including the failure at time $t_{(j)}$
$R^{*C}_{(j)}$	The subset of $R^*_{(j)}$ consisting only of cases
$R^{*NC}_{(j)}$	The subset of $R^*_{(j)}$ consisting only of non-cases
$N_{(j)}$	Size of $R_{(j)}$ (and similar for other risk sets)
N	Number of observations in the full cohort prior to any failures
N^C	Number of cases in the full cohort prior to any failures
N^{NC}	Number of non-cases in the full cohort prior to any failures
N^{SC}	Number of observations in the subcohort prior to any failures
N^{*C}	Number of cases in the subcohort prior to any failures
N^{*NC}	Number of non-cases in the subcohort prior to any failures
α	The overall subcohort sampling fraction

$\alpha_{(j)}$	The subcohort sampling fraction for $R_{(j)}$ (and similar for other risk sets)
p^C	The proportion of cases in the full cohort.
D	Indicator variable taking value 1 for cases and 0 for non-cases
Z	Vector of covariates
β	Vector of coefficients
γ	Individual coefficients
$Z_{[j]}$	Vector of covariates for the subject failing at time $t_{(j)}$
$w_{[j]}$	Weights for the subject failing at time $t_{(j)}$
$w^k_{(j)}$	Weights for the risk set $R^k_{(j)}$
Δ	Change in hazard ratio between the 25th and 75th percentiles of survival times
Z_ϕ	Vector of covariates which interact with time
β_ϕ	Coefficient vector for Z_ϕ
ϕ	Vector of coefficients for the interation of Z_ϕ with time
$h(t)$	Hazard function
$h_0(t)$	Baseline hazard function
$H(t)$	Cumulative hazard function
$H_0(t)$	Cumulative baseline hazard function
$S(t)$	Survival function
$S_0(t)$	Baseline survival function
L	Likelihood
pL	Partial likelihood
pL^*	Pseudopartial likelihood
rcs	Cox Snell residual
rmg	Martingale residual

$rs_{[j]k}$	Schoenfeld residual for covariate k failing at time $t_{(j)}$
$rsc_{[j]k}$	Scaled Schoenfeld residual for covariate k failing at time $t_{(j)}$
DE	Design effects matrix (dLR, $dBIC$)
ΔE	Design effects matrix ($dAIC$)
I	Fisher information
$\mathcal{I}$	Observed information matrix
V	The variance matrix for the vector of coefficients β
$\hat{V}$	Design-based variance estimate of V
$\hat{V}^n$	Naïeve variance estimate of V
W_D	The design-based Wald statistic

Abbreviations

All abbreviations will be introduced in the text. A summary of abbreviations follows for reference:

AIC	Akaike's information crtierion
dAIC	Modification of Akaike's information crtierion (Lumley & Scott)
BIC	Bayesian information criterion
dBIC	Modification of the Bayesian information criterion (Lumley & Scott)
pBIC	Modification of the Bayesian information criterion (Xue)
BMI	Body mass index
df	Degrees of freedom
DGM	Data-generating mechanism
ESE	Empirical standard error
FC	Full Cohort
HR	Hazard ratio
IPW	Inverse probability weighting
LOWESS	Locally Weighted Scatterplot Smoothing
lpoly	Kernel-Weighted Local Polynomial Smoothing
LR	Likelihood ratio
dLR	Naïeve modification of the Likelihood ratio (Lumley & Scott)
pLR	Modification of the Likelihood ratio (naïeve)
MCSE	Monte Carlo standard error
MSE	Mean squared error

NPH	Non-proportional hazards
PH	Proportional hazards
WC	Waist circumference
WHR	Waist-to-hip ratio
WHtR	Waist-to-height ratio
SD	Standard deviation
SE	Standard error
SC	Subcohort

Chapter 1

Introduction

1.1 Motivation

1.1.1 The Case-Cohort Design

Survival analysis concerns analysis of data where time until a specified event is the outcome of interest. Research questions in this field often take the form "How do certain characteristics affect the risk of an event occurring?". A straightforward approach is a cohort study: a cohort is defined consisting of subjects who have yet to experience the event, covariates of interest are measured for each subject, and times of events (cases) over a specified period of time are recorded. However, where the event of interest is uncommon, or covariates of interest are expensive, time-intensive or otherwise difficult to measure in large numbers, analysis of a full cohort may not be achievable. Since most of the information is contained in the cases, alternatives to analysis of the cohort often consist of designs that include all cases and a subsample of the non-cases, allowing for a substantial reduction in number of subjects and associated costs, with only a minor reduction in efficiency. The case-cohort study design is nested within a cohort study; a subcohort is randomly selected from the full cohort, covariates of interest are measured for each subcohort subject, and times of any events in the subcohort over a specified period of time are recorded. Times of events and covariate measurements are also recorded in cases outside the subcohort. The case-cohort dataset consists of all cases in the full cohort, and the non-case members of the subcohort. Compared to other alternatives to analysis of the full cohort, such as nested case-control studies, advantages of the case-cohort design include that as a random sample of the full cohort, the subcohort can be used for multiple events of interest. It can also be used to assess distributions of covariates in the population, as may be of interest in genomic studies. Further, while samples may be *collected* from the full cohort, *measurement* of potentially expensive biomarkers is required only for the cases and subcohort non-cases.

1.1.2 Post-Estimation in the Case-Cohort Design

Extant literature on the methodology of analysis of the case-cohort design is in general concerned with estimation of coefficients. In order to fully exploit the case-cohort design, we must look beyond parameter estimation. While parameter estimates are necessary for investigation of the research question posed above, they are not sufficient; proper interpretation and application of these estimates requires the ability to investigate violations of model assumptions and to compare alternative selections of explanatory covariates. Where model assumptions are violated, parameter estimates are invalid, and may lead to erroneous conclusions. While clinical judgment and other information can give indications of potential explanatory variables, model and variable selection methods can help to refine such indications. For example, clinical judgment may indicate that obesity is likely to affect risk of an event occurring, but model and variable selection methods can help in choosing between measures of obesity such as body mass index, waist circumference and waist-to-hip ratio for inclusion in the model.

The overarching aims of this thesis are to investigate post-estimation methods in the case-cohort design, to adapt existing methods or devise new methods where existing methods are inappropriate, and provide guidance for the use of these methods.

1.1.3 The Cox Proportional Hazards Model & Case-Cohort

The Cox proportional hazards (PH) model (Cox, 1972) is commonly used in the analysis of case-cohort studies. In a review of 32 papers reporting case-cohort studies published between January 2010 and March 2013, only one paper did not use some form of Cox regression in the analysis (Sharp et al., 2014). Hence, in this thesis, I consider estimation and post-estimation procedures for analysis of case-cohort studies under the Cox PH model. In this section, I outline detection of violations of model assumptions and methods of model and variable selection in the Cox PH model in the full cohort, and then describe the particular challenges posed by the case cohort design.

1.1.3.1 Full Cohort

The Cox PH model makes three key assumptions: (1) that covariates are multiplicatively related to the hazard i.e. the hazards are proportional over time; (2) that the functional form of each covariate included in the model has a linear relationship with the hazard; and (3) that the link function is exponential, i.e. the relationship between the baseline hazard function and the linear predictor is log-linear. In the full cohort, assumptions (2) and (3) can be assessed by visual inspection of smooths

of Martingale residuals against covariates and the linear predictor, respectively (Therneau et al., 1990). Methods for assessment of the proportional hazards assumption include tests of a non-zero slope in generalized linear regression models of Schoenfeld residuals or scaled Schoenfeld residuals against a function of event time, and inclusion in the model of interactions of covariates with functions of time.

In the full cohort, variable and model selection in maximum likelihood methods has a number of extant and commonly used methods, including Akaike's information criterion (AIC), Bayesian information criterion (BIC), and for nested models, Likelihood ratio and Wald tests. Apart from Wald tests, each of these methods includes the likelihood in its formula. In the full cohort, the Cox PH Model maximises a partial likelihood function rather than a likelihood function. Cox (1975) showed that large-sample properties and tests that are valid for maximum likelihood methods and an asymptotic chi-squared distribution are justified in the case where there is a partial likelihood, under broad conditions.

1.1.3.2 Case Cohort

In the case-cohort design, cases are over-represented in the case-cohort sample, and the Cox proportional hazards model must be weighted (described in Section 1.2.2), resulting in the maximisation of a pseudopartial likelihood rather than a partial likelihood.

Use of martingale residuals to detect violations of model assumptions will require adjustments to reflect the over-representation of cases. For detection of non-proportional hazards in the case-cohort design, inclusion of an interaction with time appears theoretically justified, as it relies on significance of parameter estimates, however, use of Schoenfeld and scaled Schoenfeld residuals may require weighting. For variable and model selection methods, robust Wald tests are theoretically justified, as they are again based upon significance of parameter estimates. However, the pseudopartial likelihood means that likelihood-based methods may not be readily transferable to the case-cohort design. Further, a number of different methods for weighting of the Cox model in the case-cohort design have been proposed and the choice of weighting system may have an impact on performance of post-estimation methods. In this thesis, I investigate these checks of model assumptions and model selection methods in the analysis of the case cohort design under the Cox model with a variety of weighting methods.

1.2 Background

1.2.1 Data Structure

1.2.1.1 Definitions of Time

Description and analysis of survival data may involve several distinct definitions of time. Consider a study investigating the effect of lifestyle factors on time to development of clinical arthritis. Subjects become at risk at origin times T_B, are observed from entry times $T_{(0)}$, and are followed until arthritis develops at event times T. Let superscript D refer to calendar time, superscript F refer to follow-up time, and superscript A refer to analysis time.

Calendar time refers to the dates of recruitment and events in the study. Hence subjects are recruited at times $T^D_{(0)}$, and events occur at T^D. Analysis time refers to the time-scale under which the data will be analysed, with entry times $T^A_{(0)}$ referring to the time period between the subject becoming at risk and the subject entering the study, and event times T^A referring to the time period between the subject becoming at risk and experiencing the event. Follow up times refer to the period of time for which each subject was under observation. Hence, $T^F_{(0)} = 0$ for all subjects, and $T^F = T^D - T^D_{(0)} = T^A - T^A_{(0)}$.

Unless otherwise specified, throughout the rest of this thesis, references to time without a superscript refers to analysis time.

1.2.1.2 Analysis Time-Scales

Survival analysis under the Cox PH model requires a well-defined origin and analysis time-scale. To illustrate the diversity of relationships between calendar time, follow-up time and analysis time, three examples of analysis time-scale are described below.

Scenario 1 - Recruitment as Origin: If subjects are considered to become at risk at their date of recruitment, the data can be analysed with a fixed entry time $T^A_{(0)} = T^F_{(0)} = T_B = 0 \ \forall i$ and failure occuring at times $T^A = T^F$. Such a time-scale choice is often appropriate for clinical trials, where a drug is administered at time of recruitment. Exit time is hence the duration for which subjects were observed.

Scenario 2 - Common Exposure as Origin: Alternatively, if subjects are considered to become at risk at some fixed date t^D_B, prior to recruitment, then analysis entry times $T^A_{(0)} = T^D_{(0)} - t^D_B$, and exit times $T^A = T^D - t^D_B$. Such a time-scale may be appropriate where the study is concerned with subjects who suffered some common

exposure, such as a chemical leak. Entry time to the study is hence the duration from exposure to recruitment, and exit time the duration from exposure to event.

Scenario 3 - Subject-Specific Origin: Further, if subjects are considered to become at risk at some subject-specific date T_B^D prior to recruitment , then analysis entry times $T_{(0)}^A = T_{(0)}^D - T_B^D$, and exit times $T^A = T_{(0)}^D - T_B^D$. Such a time-scale may be appropriate where subjects are considered to become at risk at birth. Entry time to the study is hence their age at recruitment, and exit time their age at event.

It is important to note that the same dataset can be analysed with differing analysis time-scales depending on the choice of the researcher.

1.2.1.3 Censoring

In the arthritis study above, the subjects were followed until arthritis developed at event times T. However, survival data commonly displays right-censoring, that is, for some subjects, the event times T are not observed. Rather, it is known only that for each censored subject, the event occurred after some censoring time T_C. Hence the recorded survival times $T_R = min(T_C, T)$. The arthritis study above could follow the subjects for a period of 10 years, with subjects who did not experience the event of interest *administratively censored* at that time.

It is also likely that it would not be possible to observe all subjects throughout the study period. Death, withdrawal of consent, relocation or other reasons may mean that subjects cannot be observed after some particular time, before experiencing the event of interest and before the end of the study. In this thesis, such subjects are denoted as lost-to-follow-up or *early censored*, with the survival time at which they were last observed referred to as the early-censoring time.

1.2.1.4 Risk Sets

Consider a dataset of ordered event times $t_{(1)}...t_{(j)}$. Subscript (0) refers to the time prior to any failure. Subscript i refers to the subject i, and subscript $[j]$ refers to the subject experiencing the event at time $t_{(j)}$. $t_{(0)i}$ refers to the entry time of subject i to the study and t_i refers to the recorded survival time of the subject i. Superscript C refers to cases and superscript NC refers to non-cases. Let N, N^C and N^{NC} refer to the number of subjects, cases and non-cases in the dataset, respectively.

For each failure time $t_{(j)}$, the risk set of observations still at risk *not including* the failure at time $t_{(j)}$ is denoted $R_{(j)}$ and is of size $N_{(j)}$ subjects, with $N^C_{(j)}$ cases and $N^{NC}_{(j)}$ non-cases.

$$R_{(j)} = i : \{t_{(0)i} < t_{(j)} < t_i\}, i \neq [j]$$

This is a non-standard definition of the risk set, as normally the subject $[j]$ failing at time $t_{(j)}$ is considered part of the risk set for that failure time. This non-standard definition is used in this thesis because, in the case-cohort design, the failure at time $t_{(j)}$ may be weighted differently to the other observations at risk at that time.

1.2.1.5 Entry and Exit Types

Both entry and exit times of an analysis dataset can be categorised as fixed or staggered, depending on whether these times are common across subjects or vary between subjects. Choice of analysis time-scale can affect whether the same data will display fixed or staggered entry and exit times. The number of subjects in each risk set $R_{(j)}$ with staggered entry and/or exit will be smaller and more variable than for similar data with fixed entry and/or exit. In terms of the effects on risk sets, times can be considered fixed even where entry and/or exit times vary between subjects, as long as all subjects enter the analysis prior to the first failure, and exit the analysis after the last failure.

In Scenario 1 above, all subjects will display the same entry time to the study. Scenario 2 will display fixed entry if subjects are recruited on the same calendar date, but if recruitment is carried out over a period of time, then analysis entry times $T^A_{(0)}$ will vary. Scenario 3 will display staggered entry unless recruitment dates were tailored so as to provide fixed entry.

If early censoring is present, the dataset will likely display staggered exit under all analysis time-scales. However, even where early censoring is not present, choice of time-scale can also effect whether the data displays staggered or fixed exit in the analysis. For Scenario 1 above, if the calendar data displayed a staggered entry, the analysis data will display fixed entry and staggered exit, as the variation in entry time has been "shifted" to the right. For Scenario 2, the type of entry and exit displayed in the calendar data will be replicated in the analysis data. For Scenario 3, it is likely the analysis data will display both staggered entry and staggered exit.

1.2.2 The Cox Model

In analysis of survival data, the hazard function $h(t)$ describes the risk of experiencing an event at time t, conditional on survival to time t, and can be considered as the number of events in the population at risk per unit time. The hazard function $h(t)$ can be considered as made up of two elements; the baseline hazard function $h_0(t)$, which describes how the risk of event per unit time changes over time at baseline levels of covariates, and the vector of coefficients β, describing how the hazard varies in response to the vector of explanatory covariates Z.

Under the Cox PH model it is assumed that, for any subject i, the ratio of the hazard over time to the hazard for any other subject j is some constant $m_{ij} = \frac{h_i(t)}{h_j(t)}$. It is also assumed that the hazard function is of the form $h(t) = h_0(t)exp(\beta^T Z)$. These assumptions allow for inferences about the ratio of the hazard between individuals without specification of the baseline hazard itself.

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)exp(\beta^T Z_i)}{h_0(t)exp(\beta^T Z_j)} = exp\{\beta^T(Z_i - Zj)\}$$

Hence, we can use estimates of β to describe the difference in risk between subjects with varying values of Z.

Using the notation and risk set structure from Section 1.2.1.4, the partial likelihood for the Cox model is given by:

$$pL(\beta) = \prod_{j=1}^{N^C} \frac{exp(\beta^T Z_{[j]})}{exp(\beta^T Z_{[j]}) + \sum\limits_{i \in R_{(j)}} exp(\beta^T Z_i)}$$

The cumulative baseline hazard function $H_0(t) = \int_0^t h_0(s)ds$, can be interpreted as the number of events that would be expected by time t, given survival to time t, for a subject with all covariates equal to 0, if the event were a repeatable process (Clark et al., 2003). Estimates of $H_0(t)$ are important for a number of post-estimation methods. In the full cohort, the Breslow estimator of $H_0(t)$ (Breslow, 1972) is given by:

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \hat{h}_0(t_{(j)}) = \sum_{t_{(j)} \leq t} \frac{1}{exp(\beta^T Z_{[j]}) + \sum\limits_{i \in R_{(j)}} exp(\beta^T Z_i)}$$

In analysis of the case-cohort design, a weighted Cox model is used, with various weighting methods proposed and used in practice. In general, weights are assigned based on some discrete classification of the subjects by categories such as subcohort status and case or non-case status. In some weighting systems, weights may also

vary by failure time $t_{(j)}$. Based on this classification, the risk set $R_{(j)}$ can hence be decomposed into K disjoint subsets, to be described later, such that $R_{(j)} = \cup_{k=1}^{K} R_{(j)}^k$. Appropriate weights $w_{(j)}^k$ can then be defined for each subset and failure time. The pseudopartial likelihood for a case-cohort sample is given by:

$$pL^*(\beta) = \prod_{j=1}^{N_{(0)}^C} \frac{w_{[j]} exp(\beta^T Z_{[j]})}{w_{[j]} exp(\beta^T Z_{[j]}) + \sum_{k=1}^{K} w_{(j)}^k \sum_{i \in R_{(j)}^k} exp(\beta^T Z_i)}$$

The weighted Breslow estimator of $H_0(t)$ is given by:

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \hat{h}_0(t_{(j)}) = \sum_{t_{(j)} \leq t} \frac{1}{w_{[j]} exp(\beta^T Z_{[j]}) + \sum_{k=1}^{K} w_{(j)}^k \sum_{i \in R_{(j)}^k} exp(\beta^T Z_i)}$$

1.3 Literature Review

In the following section I review the extant literature concerning the analysis of the case-cohort design under the Cox proportional hazards model. Literature regarding parametric modeling, accelerated failure time models, repeated events, bootstrap, and comparison of case-cohort methods with non-case-cohort alternatives is beyond the scope of this thesis and is not included.

1.3.1 Estimation of Coefficients

The vast majority of methodological papers concerning the case-cohort design under the Cox PH model concern estimation of coefficients. A number of papers have compared estimation methods for the Cox PH model in the case-cohort design. These studies and their results are described in 3.2.1.5.

1.3.1.1 Prentice Weighting

Prentice (1986) proposed a design which involves members of a subcohort, randomly selected without regard to eventual failure status, and any additional non-subchort cases. This case-cohort design is similar to that of Kupper et al. (1975) and Miettinen (1982). In Prentice's method, subcohort observations have weight 1 when at risk, and non-subcohort cases have weight 1 at their failure time and weight 0 at all other times. The asymptotic normality of estimates under Prentice weighting is shown in Self and Prentice (1988).

1.3.1.2 Self & Prentice Weighting

Self and Prentice (1988) proposed a modification to Prentice weighting. In this weighting system, subcohort observations have weights as in Prentice and non-subcohort cases have weight 0 in the denominator at all times and weight 1 in the numerator at their failure time.

1.3.1.3 Barlow Weighting

Barlow (1994) proposed an alternative weighting system, where non-subcohort cases are weighted as in Prentice (1986), subcohort cases take weight 1 at their failure time, and both subcohort non-cases and subcohort cases prior to their failure time take weight equal to the ratio of the number of cohort members at risk to the number of subcohort members at risk. These weights are hence dependent on the size of the full cohort and subcohort risk sets at each failure time $t_{(j)}$. Barlow et al. (1999) approximated these time-dependent weights by the inverse of the overall subcohort sampling fraction (the subcohort sampling fraction at time $t_{(0)}$).

1.3.1.4 Inverse Probability Weighting

Kalbfleisch and Lawless (1988) proposed an Inverse probability weighting (IPW) method for the case-cohort design. In this method, all cases take weight 1, and subcohort non-cases take weight equal to the inverse of the subcohort sampling probability.

Kulich and Lin (2004) established the asymptotic properties of this estimator and proposed a class of weighted estimators with general time-varying weights. Kang and Cai (2009) extended this estimator to studies with multivariate failure time outcomes, and Kim et al. (2013) also applied and adapted the estimator to multivariate failure time outcomes.

Chen and Lo (1999) defined a similar estimator to Kalbfleisch & Lawless, with the exception that subcohort non-cases take weight equal to the inverse of the non-case sampling fraction. The consistency and asymptotic normality of this estimator, under certain regularity conditions, was also detailed.

Chen (2001, 2004) proposed IPW methods for the case-cohort design that estimate contribution from unselected controls. This is accomplished by incorporating averages of covariates from subjects with similar failure times into calculation of weights.

1.3.1.5 Full Likelihood Approaches

The above weighting methods all result in a pseudopartial likelihood. More recently, methods considering the case-cohort design as a missing data problem have been proposed, resulting in a full likelihood approach where the likelihood expression is constructed for the complete cohort. Scheike and Martinussen (2004) proposed using the expectation maximisation (EM) algorithm for parameter estimation, whereas Kulathinal and Arjas (2006) proposed Bayesian data augmentation. Efficiency gain from the full likelihood approach is minor in case of a rare disease (Scheike and Martinussen, 2004) and the large amount of missing covariate data generated results in more computational demands. Additionally, Saarela and Kulathinal (2007) proposed another likelihood based approach where only the case-cohort data is used, but the likelihood is conditioned on the inclusion in the case-cohort sample.

1.3.1.6 Stratified Case-Cohort

The case-cohort design can be extended to include stratification. In this design, the full cohort is divided into non-overlapping sections or strata, and the subcohort is selected by stratified random sampling.

Borgan et al. (2000) considered several types of weights under case-cohort designs where the subcohort is selected by stratified random sampling. Borgan I weights are Self & Prentice weighting with additional weights applied to each subject equal to the sampling fraction for the appropriate stratum; static Borgan II weights are the application of the IPW weighting in Chen and Lo (1999), with non-case sampling fraction calculated individually for each stratum, and Borgan III weights are a score unbiased adaptation of Borgan I weights where, if the case failing at time $t_{(j)}$ is a non-subcohort case, it is included in the risk set in the place of a randomly selected subcohort member of that stratum.

Further, they propose adaptations for these methods where the above weights are replaced by their time-dependent equivalents. For example, the Borgan I weights at failure time $t_{(j)}$ are replaced by the stratum sampling fraction at that time, i.e. the ratio of the number of observations at risk in the full cohort stratum at time $t_{(j)}$ to the number of observations at risk in the subcohort stratum at time $t_{(j)}$.

Samuelsen et al. (2007) considered the application of a post-stratification approach to weighting of cohort sampling designs, including the case-cohort design, by which the methods of Chen (2001) can be placed into the framework of Borgan et al. in stratified case-cohort designs.

1.3.1.7 Estimation of Coefficient Sampling Variance

Analysis of the case-cohort design with the weighted Cox PH model requires particular care in estimation of coefficient sampling variance. Naïeve variance estimation via the standard methods for the Cox regression model are invalid, as the case-cohort sampling introduces a covariance between score terms, resulting in a larger variance for coefficient estimates than would result from standard methods.

Several variance estimators have been proposed for the case-cohort design including asymptotic variance estimators (Prentice, 1986; Self and Prentice, 1988), approximate jackknife variance estimators (Barlow, 1994; Lin et al., 1993), design-based variance estimators (Binder, 1992; Lin, 2000), super-population variance estimators (Wacholder et al., 1989), and bootstrap variance estimators (Wacholder et al., 1989). In addition, the robust variance estimator of Lin and Wei (1989) was shown by Barlow (1994) to be equivalent to a jackknife variance estimator and hence applicable to the case-cohort design.

Standard implementation of "robust" variance estimates in STATA and R statistical packages uses the Huber sandwich estimator, also known as White's estimate, the Horvitz-Thompson estimate, the working independence variance, the infinitesimal jackknife, and the Wei, Lin, Weissfeld (WLW) estimate. In this thesis, the Huber sandwich estimator is used throughout as a variance estimator that accounts for the case-cohort design. Particularly, the Huber sandwich estimator is used in Chapter 6 as the design-based variance estimator required for implementation of the modified methods for model selection described therein.

1.3.2 Estimation of Cumulative Baseline Hazard

Prentice (1986) further proposes a case-cohort estimator of cumulative baseline hazard; a weighted version of the Breslow estimator, where the case failing at time $t_{(j)}$ and the subcohort observations at risk are weighted by the inverse of the subcohort sampling fraction. The asymptotic normality of this estimator is shown in Self and Prentice (1988).

In addition to proposing a class of weighted estimators for coefficients, Kulich and Lin (2004) also propose related estimators for the cumulative baseline hazard. However, to my knowledge, performance of estimation of cumulative baseline hazard under IPW in the case-cohort design has not been investigated.

1.3.3 Post-Estimation

Methodological literature on post-estimation procedures of case-cohort data under the Cox PH model is sparse.

To the best of my knowledge, detection of inappropriate functional form and inappropriate link function have not been studied in the case-cohort design, and detection of non-proportional hazards has been investigated only by Xue et al. (2013), who define "case-cohort Schoenfeld residuals" under Prentice weighting, and assess the use of the correlation of these residuals with functions of event time in detection of non-proportional hazards.

Barlow (1997) proposes a modification of the Pettitt and Bin Daud (1989) Likelihood Displacement measure of individual influence to allow for multiple failure time data and case-cohort designs. To the best of my knowledge, use of deviance residuals in the case-cohort design has not been studied.

Ganna et al. (2012) consider risk prediction measures in the case-cohort design, including the Grønnesby and Borgan (1996) goodness-of-fit test for calibration, the net reclassification improvement, and the concordance-index, and conclude that case-cohort designs can be used in evaluation of the prediction ability of new markers.

For model selection, Lumley and Scott (2013) consider the case-cohort design as a special case of complex survey sampling in a paper where they introduce a modified Likelihood Ratio test dLR.

In complex survey sampling, Xu et al. (2013), propose replacement of the log maximum likelihood with the log maximum pseudolikelihood for a modified BIC ($pBIC$), and Lumley and Scott (2015) build on their 2013 paper to describe modifications to AIC and BIC ($dAIC$ and $dBIC$), however, these papers do not refer to the case-cohort design specifically.

Newcombe et al. (2018) propose use of Bayesian variable selection, a method based on Bayesian sparse logistic regression, and compare its performance with (a) one-at-a-time significance testing of potential variables, and (b) forwards stepwise selection. Ni et al. (2016) propose use of the smoothly clipped absolute deviation penalty, a penalty-based variable selection procedure.

1.3.4 The Case-Cohort Design in Practice

Sharp et al. (2014) conduct a review of 32 papers reporting case-cohort studies published from Jan 2010 to March 2013. Some of these used case cohort samples from centres or groups of centres from the European Prospective Investigation into Cancer and Nutrition (EPIC) study, a cohort study with 521,000 subjects recruited across 23 centres in 10 European countries (Riboli et al., 2002). Treating each EPIC centre/group of centres as a separate cohort, the 32 papers were based on 17 cohorts.

Nine of the 17 original cohorts used stratified sampling to select the subcohort. The stratifying variables were age, gender, race, location or a combination of these. The median size of the full cohort before exclusions was 48,532 (interquartile range 14,610 to 124,426). The median subcohort sampling fraction before exclusions was 4.1% (interquartile range 3.7% to 9.1%)

As expected, the published studies were carried out on the case-cohort samples after some exclusion of observations, with exclusions dependant on the purpose of the study. Complete information on full cohort size, subcohort size and number of cases *post-exclusions* was not available for 8 of the 32 papers detailed in the report. This complete information post-exclusions for the remaining 24 papers is summarized in Table 1.1 below (pers comm Sharp).

Table 1.1: Summary Statistics (Post-Exclusions) for 24 Case-Cohort Papers

	min	25th pctile	median	75th pctile	max
Full Cohort N	950	9,630	27,548	76,364	340,234
Full Cohort Case N	77	421	597	2,007	12,403
Subcohort N	190	1,290	1,920	3,034	16,154
Full Cohort Case Percentage	0.6	2.4	3.6	5.3	24.1
Subcohort Sampling Fraction	0.029	0.044	0.065	0.131	0.255
Case-Cohort Non-Case to Case Ratio*	0.70	1.26	1.98	3.23	6.17

*calculated assuming equal case and non-case exclusion from the subcohort

Of the 32 papers, a single paper used logistic regression, with the remainder using some form of the Cox PH model. In the papers using Cox regression, ten used unweighted Cox regression, which is inappropriate for the case-cohort design, ten used Prentice weights, seven used Barlow weights, three papers were unclear as to which weights had been used, and only one paper used a full-likelihood approach. Of the 31 papers using Cox regression, 12 reported that the proportional hazards assumption was tested. Nine used age as the underlying timescale rather than study duration. Seventeen papers specified that robust standard errors were calculated.

1.4 Research Focus

The literature review reveals a number of areas where further investigation could allow greater exploitation of the case-cohort design in the Cox PH model.

Despite a number of alternative methods for estimation of coefficients having been proposed, Prentice and Barlow weighting methods appear to be those most commonly used in practice. In this thesis, I consider Prentice, Barlow, and IPW methods, with IPW included due to reports from the literature that its performance is superior to Barlow (see Section 3.2.1.5), its ease of implementation using standard statistical software, and its possible value in post-estimation methods where residuals and other quantities must be weighted in the case-cohort design.

In the majority of literature considering methods for the case-cohort design, simulation studies generally consist of full cohort sizes, sampling fractions, and case percentages that are dissimilar to those seen in practice. That is, full cohort sizes are small, and sampling fractions and case percentages are high (see Section 3.2.1.5.1, where simulation studies from the literature are described). In this thesis the design of the simulation studies aims to provide circumstances which are more similar to those encountered when using case-cohort designs in practice.

There is a clear dearth of research in post-estimation methods in the case-cohort design. In this thesis, I will assess application and modification of full cohort methods for detection of violation of Cox PH model assumptions in the case-cohort design. Further, I will investigate the application and modification of standard model and variable selection methods for use in the case-cohort design.

Finally, in order for methods to be useful in practice, they should be possible to implement in statistical software without overly complex coding and with minimal opportunities for user error. In an appendix I will include sample STATA code for these methods and comment on the the practicality of their implementation.

1.5 Dissertation Structure

In this introductory chapter, I have outlined the motivation for the investigation of the Cox PH model in the case cohort design, with regard to estimation and post-estimation procedures. Further, I have outlined the main considerations of data structure in this investigation, specifically, censoring, time-scale, and risk sets, and defined a case-cohort pseudo-partial likelihood for the weighted Cox model. I have

reviewed the relevant extant literature, and described my research focus.

In Chapter 2, I describe and justify a general data-generating mechanism for the simulation studies in this thesis. Specific simulation studies have any alterations to this general data generating mechanism described in the relevant chapter.

In Chapter 3, I investigate estimation of coefficients and cumulative baseline hazard in the case-cohort design. Simulation studies are performed to compare Prentice, Barlow, and IPW methods. I also investigate whether post-stratification on case or non-case status and failure time $t_(j)$ provide improvements in performance.

In Chapter 4, I investigate use of martingale residuals in detection of inappropriate functional form in the case-cohort design. A simulation study is performed with statistical assessment of non-linearity of weighted linear splines used as a proxy for subjective visual assessment of weighted smooths.

In Chapter 5, I investigate detection of non-proportional hazards in the case-cohort design. Methods involving Schoenfeld residuals, scaled Schoenfeld residuals, and inclusion of an interaction with time are assessed by a simulation study.

In Chapter 6, I investigate methods for model and variable selection in the case-cohort design. The robust Wald test, and modifications to the Likelihood Ratio test, Bayesian Information Criterion, and Akaike Information Criterion are assessed by simulation study. Modifications to likelihood-based methods include naïeve replacement of the partial likelihood with the pseudopartial likelihood, the modified Likelihood Ratio test as proposed by Lumley & Scott as applicable to the case-cohort design, and application to the case-cohort design of the modified AIC and BIC proposed by Lumley & Scott for complex survey data.

In Chapter 7, I apply the methods described in the previous chapters to a real-world dataset, InterAct; a case–cohort study designed to allow for examination of genetic and lifestyle factors on incidence of type 2 diabetes in the EPIC Study.

In Chapter 8, I summarize the conclusions of this thesis and discuss possible avenues for future research.

Finally, in an Appendix, provide sample STATA code for the methods described in this thesis. I also discuss the general approaches and challenges to implementation of the methods under Prentice and IPW weighting in STATA.

Chapter 2

General Data-Generating Mechanism

2.1 Introduction

In this thesis, a number of different simulation studies are presented. The overall goal of each simulation study was to assess the performance of case-cohort estimation and post-estimation procedures, often with regard to a "gold standard" such as full-cohort results or a commonly-used method. Methods were assessed across different combinations of case percentages, non-case to case ratios, β, time-scales, and other factors to assess whether the performance of methods is affected by such factors. In designing these simulation studies, I aimed to provide circumstances which are more similar to those encountered when using case-cohort designs in practice than are sometimes found in the literature, such as in the studies described in Section 3.2.1.5.1. Further, I aimed to provide a level of consistency in follow-up times, case percentages and subcohort sampling fractions, to mitigate any potential confounding effects that large amounts of variation in these factors may cause when I assessed the performance of methods across the combinations of factors listed above. A general data-generating mechanism applies to all simulation studies in this thesis, unless otherwise specified. In this chapter, I describe and justify this general data-generating mechanism. Choice of covariates, covariate functional forms, associated coefficients, and other model parameters vary according to the purpose of the simulation study and are detailed separately in each chapter.

In Section 2.2 I describe my use of the Cambridge centre of the InterAct dataset to model plausible baseline survival distribution for events such as would be the focus of a case-cohort study. In Section 2.3.1 I describe the methods of Bender et al. (2005) for simulation of survival times with proportional hazards under fixed

entry and my adaptation to allow for simulation of survival times under staggered entry. In Section 2.3.1.1 I describe how scaling can be applied to the parameters obtained in Section 2.2 to obtain broadly similar follow-up times across simulation studies, regardless of differing hazard ratios and case percentages. In Section 2.3.2 I describe the methods of Bender et al. (2005) for simulation of survival times with non-proportional hazards under fixed entry and my adaptation to allow for simulation of survival times under staggered entry. In Section 2.3.2.1 I describe how, when simulating datasets with non-proportional hazards, the direction and magnitude of the non-proportionality of the hazard can be specified.

In Section 2.4 I justify my choices of subcohort size, sampling fractions, and case percentages, describe the procedure I use to administratively censor the dataset to achieve a desired case percentage, and describe the sampling procedure by which I achieve a desired subcohort sampling fraction and the final case-cohort dataset. In Section 2.7 I provide summary statistics for full cohorts generated via this data-generating mechanism, and finally, in Section 2.8 I discuss the potential limitations of this general data-generating mechanism, and alternatives that were considered.

2.2 Baseline Survival Distributions

In this thesis, a variety of estimation and post-estimation procedures were assessed for a variety of covariates and coefficients. This required baseline survival distributions from which event times and early censoring times could be simulated. In order that the simulation studies in this thesis were reflective of datasets that might be seen in practice, initial parameters for simulation of survival times were modelled from the Cambridge centre of the InterAct dataset, a case-cohort study described briefly below and in more detail in 7.2. Note that the aim of this procedure is not to simulate the Cambridge centre of the InterAct dataset itself, but rather to obtain plausible baseline survival distributions for events such as would be the focus of a case-cohort study.

2.2.1 The InterAct Study

The InterAct case-cohort study is designed to allow for examination of genetic and lifestyle factors on incidence of type 2 diabetes in the EPIC Study. Standard anthropometric data and biological samples were collected from 346,055 of 455,680 individuals over 11 study locations. Individuals with prevalent diabetes (n=5821) at baseline were excluded. The InterAct study consists of 12,403 incident type 2 diabetes cases and a randomly selected subcohort of 16,154 individuals, drawn from

a total cohort of 340,234 participants with 3.99 million person-years of follow-up. The participants, methods, study design and measurements are described in more detail in Chapter 7 and comprehensively in InterAct Consortium et al. (2011).

The Cambridge centre of the InterAct dataset consists of 960 subcohort non-cases, 29 subcohort cases, and 758 non-subcohort cases drawn from a full cohort of 23,081 subjects. Following exclusion of 77 subcohort non-cases, 3 subcohort cases and 41 non-subcohort cases with missing data for Physical Activity, the Cambridge centre of the InterAct dataset consists of 1,626 subjects, comprising 743 non-subcohort cases, 26 subcohort cases and 883 subcohort non-cases, drawn from a full cohort of 20,023. The subcohort sampling fraction is 4.38%, the subcohort non-case sampling fraction is 4.41%, and the full-cohort case percentage is 3.58%. For study end defined as 31st December 2007, 121 subjects were lost-to-follow-up, comprising 13.31% of the subcohort sample and 13.7% of the non-cases. In the rest of this section, for brevity, the Cambridge centre of the InterAct dataset after exclusions on Physical Activity is referred to as the InterAct dataset.

The median and interquartile range for survival times of cases under duration as time-scale and age as time-scale in the InterAct dataset is shown in Table 2.1.

Table 2.1: Survival Times of Cases (Years) - InterAct

	Min	25th Percentile	Median	75th Percentile	Max
Follow-Up Time	0.25	4.73	6.14	7.66	12.69
Age at Event	43.09	62.69	69.25	74.98	85.56

2.2.2 Modeling of Full Cohort Parameters from InterAct

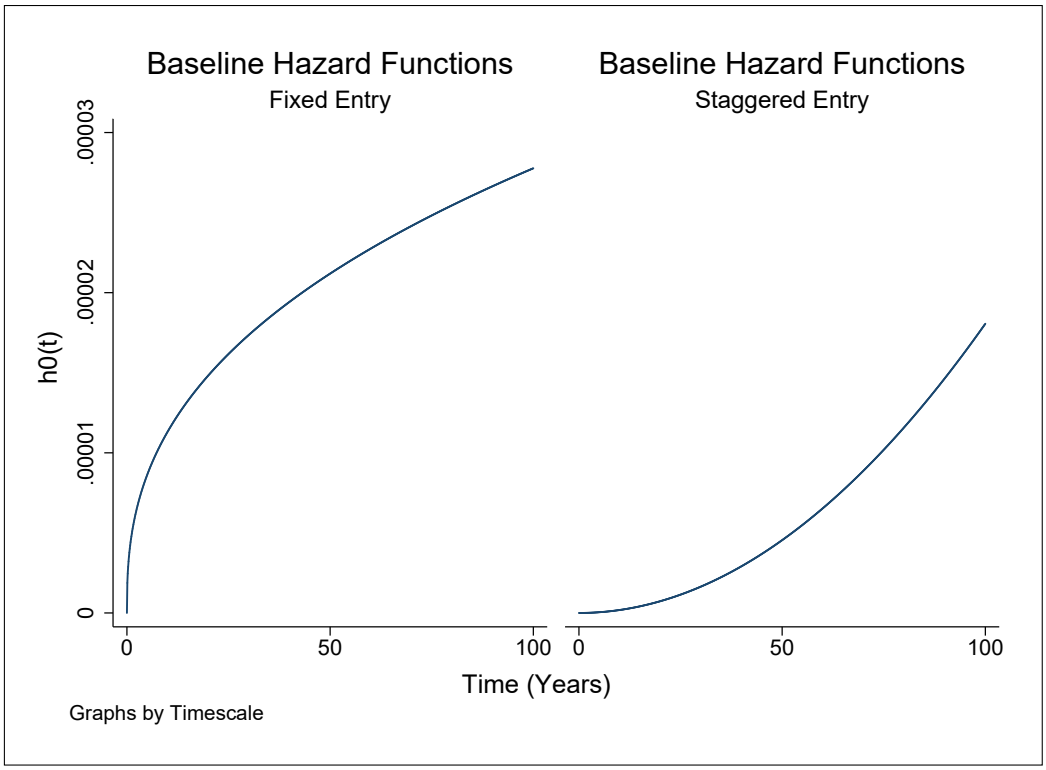
In this thesis I simulate survival times using two distinct time-scales, each modelled from the InterAct dataset using a different analysis time-scale. In the first, denoted fixed entry, subjects become at risk at recruitment to the study, such that entry times $T_{(0)} = 0 \ \forall i$. In the second, denoted staggered entry, subjects become at risk at "birth", such that entry times $t_{(0)i}$ vary by subject. Note that regardless of the "true" type of entry, a researcher could choose to analyse the data using study duration or using age as time-scale, and indeed, the same dataset may be analysed using different timescales as I do below. However, in simulation studies in this thesis, analysis time-scale corresponds to simulation time-scale.

To obtain a baseline survival distribution for an event of interest, a selection of covariates were centred with mean equal 0, so as to allow for parametric modeling of a baseline survival distribution for an "average" subject. The covariates included in the model were: Physical Activity, a 4-level categorical covariate, treated here as continuous; body mass index (BMI), a continuous covariate; Sex, a binary covariate; Any Smoking History, a binary covariate; Any Hypertension, a binary covariate; Any Hyperlipidemia, a binary covariate; and Family History of Diabetes, a binary covariate. The event of interest was diagnosis of type 2 diabetes.

The data was modelled separately using study duration as time-scale and age as time-scale. For duration as time-scale, entry time was set as 0, and exit times were calculated as days since recruitment. For age as time-scale, entry and exit times were calculated as age in days at recruitment and age in daysat event or censoring. For each time-scale, a parametric Weibull model was fitted for the mean-centred covariates, with IPW Classic weighting, where cases had weight 1 and non-cases were weighted by the inverse of the non-case sampling fraction.

The resulting parameters for baseline Weibull distributions were, for time in years, $\lambda = 6.32 \times 10^{-10}$, $v = 2.99$, and $\lambda = 4.17 \times 10^{-2}$, $v = 1.39$, for data modelled with age as time-scale and duration as time-scale, respectively.

Figure 2.1: Baseline Hazard Functions as Modelled from InterAct for Type 2 Diabetes as Outcome



2.3 Simulation of Survival Times

Bender et al. (2005) describe methods for simulation of survival times T for simulation studies regarding Cox proportional hazards models. These methods are detailed for simulations of survival times for covariates with proportional hazards and non-proportional hazards, and with baseline survival distributions under the exponential, the Weibull and the Gompertz models. In the general form $T = H^{-1}(-log(U))$, where $U \sim Uniform(0, 1)$.

A limitation of these methods as described is that only fixed entry is specified. I have extended these methods for a Weibull model with staggered entry and detail these methods below under proportional hazards and non-proportional hazards.

In this thesis, for staggered entry, values for age at recruitment $t_{(0)i}$ are drawn using $Normal(60, 10)$, truncated at 40 and 80 years. For fixed entry, $t_{(0)i} = 0 \ \forall i$

2.3.1 Survival Times under Proportional Hazards

Let Z be a vector of covariates with associated coefficient vector β.

For a Weibull distribution with entry times $t_{(0)i}$, scale parameter λ, shape parameter v, and $h(t) = \lambda v t^{v-1} exp(\beta^T Z)$:

$$H_0(t, t_{(0)}) = \int_{t_{(0)}}^{t} h_0(s)ds = \int_{t_{(0)}}^{t} \lambda v s^{v-1} ds = \lambda t^v - \lambda t_{(0)}^v$$

$$H_0^{-1}(t) = \left(t_{(0)}^v + \frac{t}{\lambda} \right)^{\frac{1}{v}}$$

$$T = \left(t_{(0)}^v - \frac{log(U)}{\lambda exp(\beta^T Z)} \right)^{\frac{1}{v}}$$

Hence, for fixed entry with $t_{(0)i} = 0 \ \forall i$:

$$T = \left(-\frac{log(U)}{\lambda exp(\beta^T Z)} \right)^{\frac{1}{v}}$$

2.3.1.1 Scaling of λ

In order to compare results from simulation studies, I aimed for the full cohorts in each simulation study to have broadly similar survival times regardless of differing hazard ratios and case percentages. This was to mitigate any confounding effects that large differences in survival times may have on interpretation of the results of

simulation studies. It is possible to obtain medians and interquartile ranges of survival times that are reasonably consistent across differing combinations of covariates and associated coefficients and case percentages by scaling the λs used to generate the survival times.

The scaling factor has two components. The first accounts for varying vectors of covariates and coefficients by ensuring that the expectation of $h(t)$ at a particular survival time remains constant regardless of the vectors of covariates and coefficients used for simulation. Hence this first component is set as $\left(\frac{1}{N}\sum exp(\beta^T Z)\right)^{-1}$. The second component accounts for varying case percentages in the data, and was found by performing a grid search with case percentages 1%, 5%, 10%, and 15%. The following procedure was performed separately for each time-scale. A lower and upper bound for the scaling factor was found by exploratory simulation. Scaling factors assessed were for 200 increments from the lower to the upper bound. For each case percentage, survival times were generated for 100 full cohort datasets of size N=10,000 with a single covariate $\sim N(0,1)$. The scaling factor resulting in a median survival time closest to that found in the Cambridge Centre of the InterAct dataset was recorded, resulting in a dataset consisting of 100 scaling factors $\times$ 4 case percentages. A linear regression of scale against case percentage was performed and the resulting coefficient used to provide the second component of the scaling factor.

Table 2.2 summarizes the parameters modeled from InterAct, the survival distributions, and the scaling derived from the above methods.

Table 2.2: Survival Distributions for Simulation Studies

Staggered Entry/Exit	
Observation Becomes at Risk	$t_0 \sim$ Normal(60,10); truncated at 40 & 80
Baseline Survival Distribution:	Weibull: $\lambda = 6.32 \times 10^{-10}$, $v = 2.99$
Baseline λ scaling:	$0.28p^C \left(\frac{1}{N}\sum exp(\beta^T Z)\right)^{-1}$

Fixed Entry	
Observation Becomes at risk	$t_0 = 0$
Baseline Survival Distribution	Weibull: $\lambda = 4.17 \times 10^{-2}$, $v = 1.39$
Baseline λ scaling:	$0.35p^C \left(\frac{1}{N}\sum exp(\beta^T Z)\right)^{-1}$

p^C = proportion of cases; N = full cohort size; t˙0 = time of entry to study

2.3.2 Survival Times Under Non-Proportional Hazards

Survival times for non-proportional hazards can be simulated by incorporating a time dependent effect ϕ. Let Z be a vector of covariates which do not interact with time, with coefficient vector β, and Z_ϕ be a vector of covariates which interact with time, with coefficient vector β_ϕ, and vector of interaction effect with log time ϕ.

For a Weibull distribution with $h(t) = \lambda v t^{v-1} exp(\beta^T Z + \beta_\phi^T Z_\phi + log(t)\phi Z_\phi)$ and entry times $t_{(0)i}$:

$$H(t) = \int_{t_{(0)}}^{t} h(s)ds = \int_{t_{(0)}}^{t} \lambda v s^{v-1} exp\left(\beta^T Z + \beta_\phi^T Z_\phi + log(s)\phi Z_\phi\right) ds$$

$$H(t) = \lambda v exp(\beta^T Z + \beta_\phi^T Z_\phi) \frac{t^{\phi Z_\phi + v} - t_{(0)}^{\phi Z_\phi + v}}{\phi Z_\phi + v}$$

$$H^{-1}(t) = \left(t \frac{\phi Z_\phi + v}{\lambda v exp(\beta^T Z + \beta_\phi^T Z_\phi)} + t_{(0)}^{\phi Z_\phi + v} \right)^{\frac{1}{\phi Z_\phi + v}}$$

$$T = \left(-log(U) \frac{\phi Z_\phi + v}{\lambda v exp(\beta^T Z + \beta_\phi^T Z_\phi)} + t_{(0)}^{\phi Z_\phi + v} \right)^{\frac{1}{\phi Z_\phi + v}}$$

Hence, for fixed entry with $t_{(0)i} = 0 \ \forall i$:

$$T = \left(-log(U) \frac{\phi Z_\phi + v}{\lambda v exp(\beta^T Z + \beta_\phi^T Z_\phi)} \right)^{\frac{1}{\phi Z_\phi + v}}$$

2.3.2.1 Choice of ϕ and β_ϕ

For simulation of datasets where covariates(s) display non-proportional hazards by the methods described in Section 2.3.2, ϕ and β_ϕ must be specified. In this thesis, ϕ and β_ϕ are chosen by the following method, which allows for stipulation of the direction and magnitude of the non-proportionality of the hazard.

ϕ and β_ϕ are chosen with regard to survival times for a reference proportional hazards dataset, where the reference dataset has the same time-scale, case percentage, vector of covariates, and vector of coefficients for those covariates with proportional hazards, as will be specified for the non-proportional hazards dataset. Let Z_k be a covariate to be simulated under non-proportional hazards, and let β_k refer to a specified coefficient of Z_k in the reference dataset. Let t_{p25}, t_{p50}, and t_{p75} refer to the survival times for the 25th percentile, median, and 75th percentile of cases in the reference dataset.

The aim is to choose ϕ_k and β_{ϕ_k} so that (1) at time t_{p50}, $\beta_{\phi_k} + \phi_k ln(t) = \beta_k$,

$$\beta_{\phi_k} + ln(t_{p50})\phi_k = \beta_k$$

$$\beta_{\phi_k} = \beta_k - ln(t_{p50})\phi_k$$

and (2) there is a specified change in hazard ratio Δ between times t_{p25} and t_{p75}.

$$exp(\beta_{\phi_k} + ln(t_{p75})\phi_k) = \Delta exp(\beta_{\phi_k} + ln(t_{p25})\phi_k)$$

$$\phi_k = \frac{ln(\Delta)}{ln(t_{p75}) - ln(t_{p25})}$$

With this method, the median survival time for a particular coefficient β_k under proportional hazards is used as a "pivot" for choice of ϕ_k and $\beta_{\phi k}$, so as to give a desired ratio between the Hazard Ratios at the 25th and 75th percentiles of survival times of cases under proportional hazards. Converging hazards, diverging hazards, and crossing hazards of different magnitudes can all be achieved by appropriate choice of Δ and β_k.

2.4 Subcohort Size, Censoring & Sampling

As described in Chapter 1, Sharp et al. (2014) conducted a review of 32 published analyses of case-cohort studies, summarized in Table 1.1. In this thesis, subcohort sizes considered are 200 and 1000, chosen based on the minimum and 25th percentile of published analyses for which complete information on full cohort size, subcohort size and number of cases post-exclusions was available. Subcohort sampling fractions considered are 3% and 15%, with full cohort case percentages such as to give non-case to case ratios of 1:1 and 4:1, both chosen to incorporate the interquartile ranges from same. The full cohort sizes and number of cases corresponding to these subcohort sizes, sampling fractions and non-case to case ratios are given in Table 2.3.

2.4.1 Right-Censoring

Following simulation of survival times as described in previous sections, the resulting dataset has a known event time for each subject. In this section, I describe the methods by which a desired full cohort case percentage is achieved. Regardless of time-scale, in this thesis I use a single study design where recruitment is considered to take place at the same calendar time for all subjects, and administrative censoring takes place after a fixed period of follow-up t^F for all subjects. Note that in practice, studies will generally display recruitment over a period of time, for logis-

tical reasons, and that "fixed" administrative censoring may also take place over a period of time, for same. Studies will also likely display a degree of loss to follow-up.

t^F is chosen to achieve a specified proportion of cases p^C. The uncensored dataset consists of N subjects, each with uncensored survival time T_U and entry time to the study $T_{(0)}$. For each subject, the uncensored follow-up time $T_U^F = T_U - T_{(0)}$. R_U = the rank order of T_U^F. The scalar t^F = the value of T_U^F for the subject with $R_U = Np^C$, and the administratively censored survival time $T_A = T_{(0)} + t^f$. The final recorded survival time $T_R = min(T_U, T_A)$.

Let the case indicator $D = 1$ if $T_R = T_U$ and $D = 0$ else. The resulting full cohort dataset consists of $N_{(0)}$ subjects with Np_C cases, final recorded survival time T_R, entry time to the study $T_{(0)}$ and binary indicator variable D indicating case or non-case status.

2.4.2 Case-Cohort Sampling

In this thesis, simulated datasets are randomly sampled so that the resulting dataset will have a specified subcohort sampling fraction α. For each observation, a random value U is independently drawn from $Uniform(0, 1)$. R_S = the rank order of U for each subject and the scalar u_S = the value of U for the subject with $R_S = N\alpha$. The subcohort indicator $S = 1$ if $U \leq u_S$ and $S = 0$ else. The final case-cohort dataset consist of subjects with $S = 1$ or $D = 1$.

Table 2.3: Full Cohort Sizes and Number of Cases Considered in this Thesis

N^{SC} : Subcohort size		200		1000	
$\frac{N^{NC}}{N^C}$ Non-case to case ratio	α (%) Sampling Fraction	N Full cohort size	N^C No. of cases	N Full cohort size	N^C No. of cases
4	15	1,333	48	6,667	241
	3	6,667	50	33,333	248
1	15	1.333	170	6,667	850
	3	6,667	194	33,333	970

2.5 Independence and Replicates

Separate full cohorts are simulated for each combination of time-scale, subcohort size, subcohort sampling fraction, non-case to case ratio, vector of covariates and vector of coefficients. Each simulation study is carried out over 1000 replicates.

2.6 Centering

In this thesis, covariates are centred with mean equal 0 prior to simulation of survival times. Covariates are also centred with mean equal 0 prior to analysis, with IPW Classic weighting used to calculate the means in the case-cohort sample.

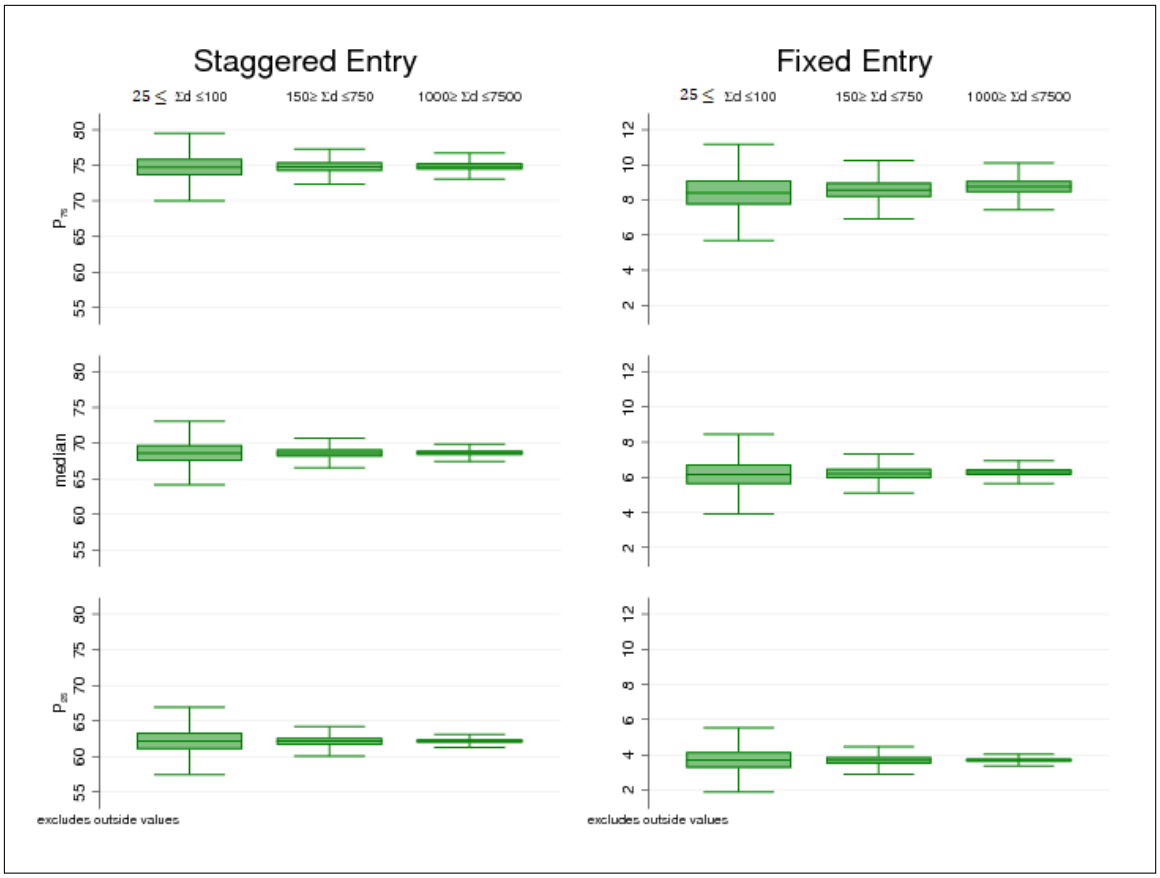
2.7 Checking the Simulation Design

In order to demonstrate consistency of survival times in datasets with proportional hazards under this general data-generating mechanism, 100 full cohort replicates were generated for the following scenarios:

- Time-Scale: staggered entry, fixed entry
- Case Percentage: 0.5%, 1%, 5%, 10%, 15%
- Full Cohort Size: 1000, 5000, 10000, 20000, 50000
- Covariates: two covariates centred with mean equal 0; both $\sim$ Binomial(0.5), both $\sim$ Normal(0, 1), one $\sim$ Binomial(0.5) and the other $\sim$ Normal(0,1)
- Correlation: covariates drawn independently, covariates with $\rho = 0.5$
- Coefficients: ln(2)/SD, ln(0.5)/SD, ln(.8)/SD, ln(1.25)/SD

Figure 2.2 shows boxplots for recorded survival times of cases by entry type and number of cases in the dataset. Staggered entry displays greater variability of survival times than fixed entry. Smaller numbers of cases in the datasets correspond to increasing variability of survival times. Survival times for each benchmark are broadly consistent, with medians similar to the survival times seen in the Cambridge centre of the Interact Dataset.

Figure 2.2: Survival Times (Years) in the General Simulation Design



$\sum d$ = number of cases

y axis starts at birth for staggered entry and at recruitment for fixed entry

To demonstrate survival times for datasets with non-proportional hazards under this general DGM, 100 full cohorts were generated for the following scenarios:

- Time-Scale: staggered entry, fixed entry
- Full Cohort: 10,000
- Case Percentage: 4%
- Covariates: two covariates centred with mean equal 0 generated independently $\sim$ Binomial(0.5).
- Coefficients (reference datasets): ln(2)/SD, ln(0.5)/SD, ln(.8)/SD and ln(1.25)/SD.
- Δ (for one covariate) = $0.5, 0.8, 1, 1.25, 2$. $\Delta = 1$, or no change, serves as a reference for the corresponding results under proportional hazards.

Table 2.4 shows values of ϕ and β_ϕ for the scenarios detailed above. Note that β_ϕ is the coefficient of the covariate Z_ϕ at time t = 1 where log(t) = 0. As time in days t increases, the influence of β_ϕ on the overall hazard, compared to that of ϕ decreases.

Hence, while β_ϕ is large in Table 2.4, the corresponding hazard ratios (HR) shown in Fig. 2.3 are less extreme. Fig. 2.4 shows boxplots for the median survival times of cases in each replicate. Divergence of HR from the reference corresponds to that of the median survival times, with increase/decrease of median survival times dependant on the signs of ϕ and β_k.

Table 2.4: ϕ and β_ϕ Corresponding to Δ and HR/SD

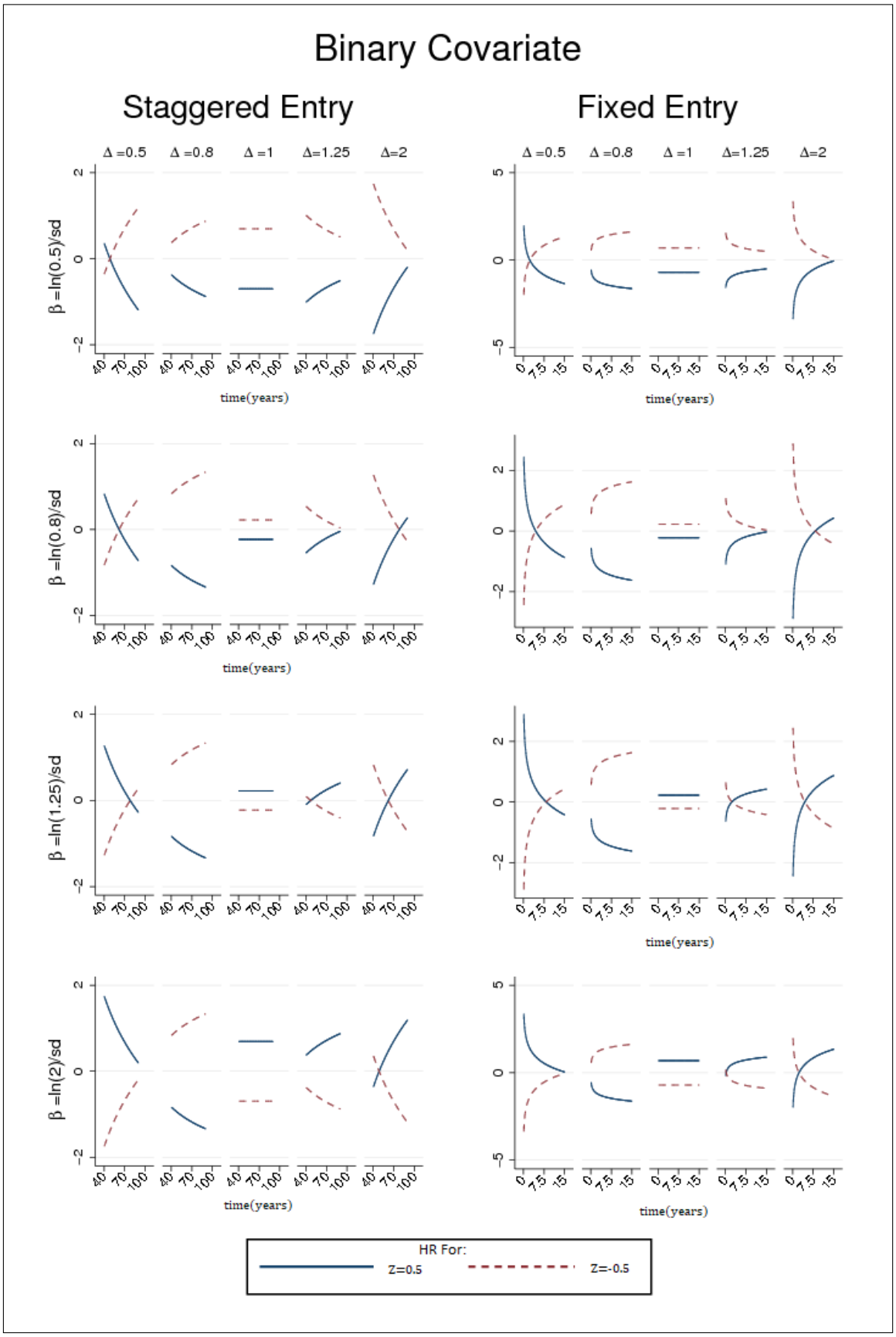
Time-Scale	Covariate	HR/SD	Δ:									
			0.5		0.8		1		1.25		2	
			ϕ	β_ϕ	ϕ	β_ϕ	ϕ	β_ϕ	ϕ	β_ϕ	ϕ	β_ϕ
Stag. Entry	Binary	0.5	-3.87	37.87	-1.25	11.25	0	-1.39	1.25	-14.02	3.87	-40.64
		0.8	-3.87	38.81	-1.25	10.33	0	-0.45	1.25	-13.08	3.87	-39.70
		1.25	-3.87	39.70	-1.25	10.33	0	0.45	1.25	-12.19	3.87	-38.81
		2	-3.87	40.64	-1.25	10.33	0	1.39	1.25	-11.25	3.87	-37.87
	Normal	0.5	-3.87	38.56	-1.25	10.33	0	-0.69	1.25	-13.33	3.87	-39.95
		0.8	-3.87	39.03	-1.25	10.33	0	-0.22	1.25	-12.86	3.87	-39.48
		1.25	-3.87	39.48	-1.25	10.33	0	0.22	1.25	-12.41	3.87	-39.03
		2	-3.87	39.95	-1.25	10.33	0	0.69	1.25	-11.94	3.87	-38.56
Fix. Entry	Binary	0.5	-1.44	9.71	-0.46	0.71	0	-1.39	0.46	-4.96	1.44	-12.48
		0.8	-1.44	10.65	-0.46	0.71	0	-0.45	0.46	-4.02	1.44	-11.54
		1.25	-1.44	11.54	-0.46	0.71	0	0.45	0.46	-3.12	1.44	-10.65
		2	-1.44	12.48	-0.46	0.71	0	1.39	0.46	-2.18	1.44	-9.71
	Normal	0.5	-1.44	10.40	-0.46	0.71	0	-0.69	0.46	-4.26	1.44	-11.79
		0.8	-1.44	10.87	-0.46	0.71	0	-0.22	0.46	-3.79	1.44	-11.32
		1.25	-1.44	11.32	-0.46	0.71	0	0.22	0.46	-3.35	1.44	-10.87
		2	-1.44	11.79	-0.46	0.71	0	0.69	0.46	-2.88	1.44	-10.40

Δ change in hazard ratio between 25th and 75th percentile of reference survival times

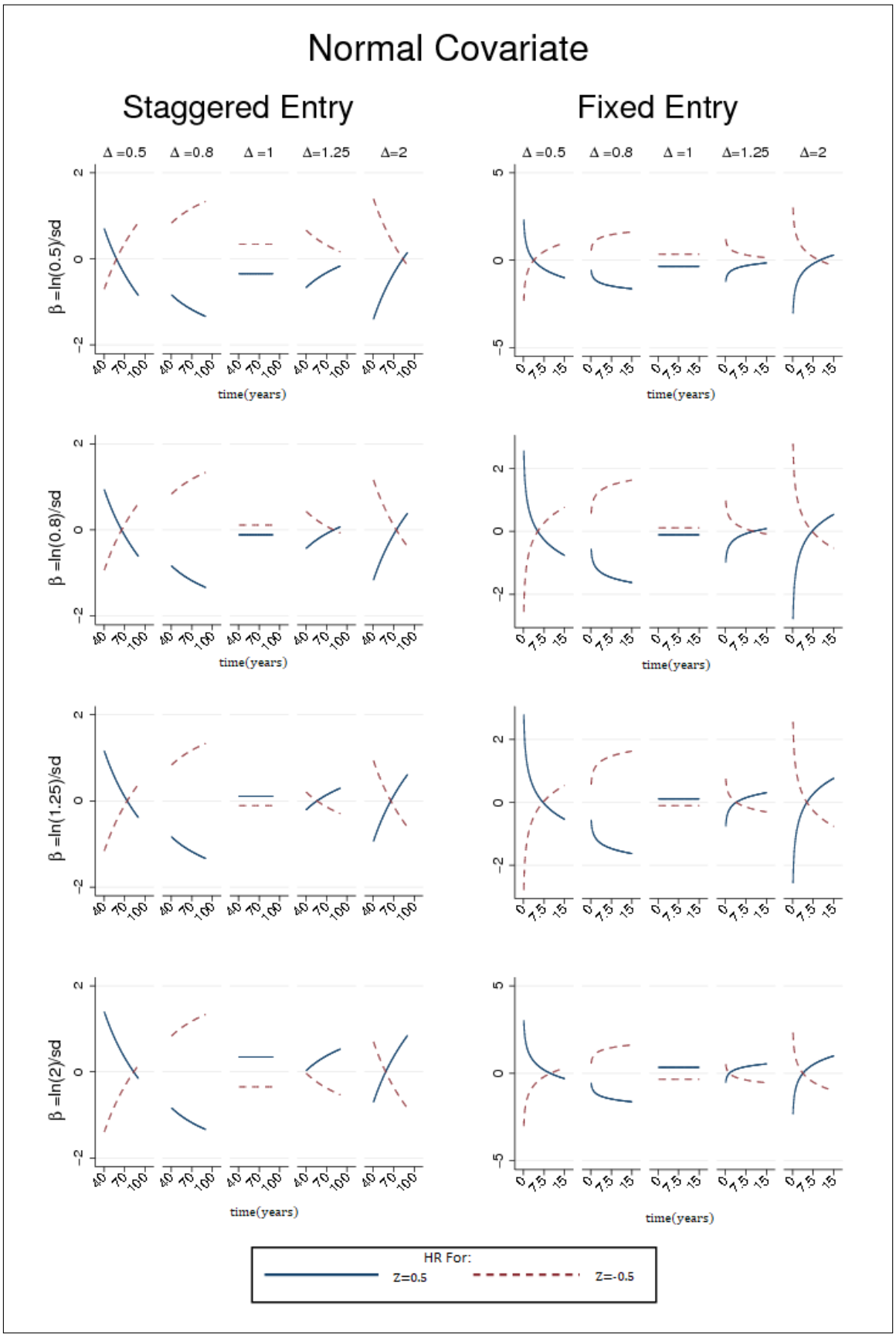
Coefficient of covariate Z_ϕ is $beta_\phi + log(t) * \phi$

HR/SD refers to the coefficient of the reference dataset

Figure 2.3: Hazard Ratios against Survival Time (Years) for varying Δ and, varying β in the reference dataset

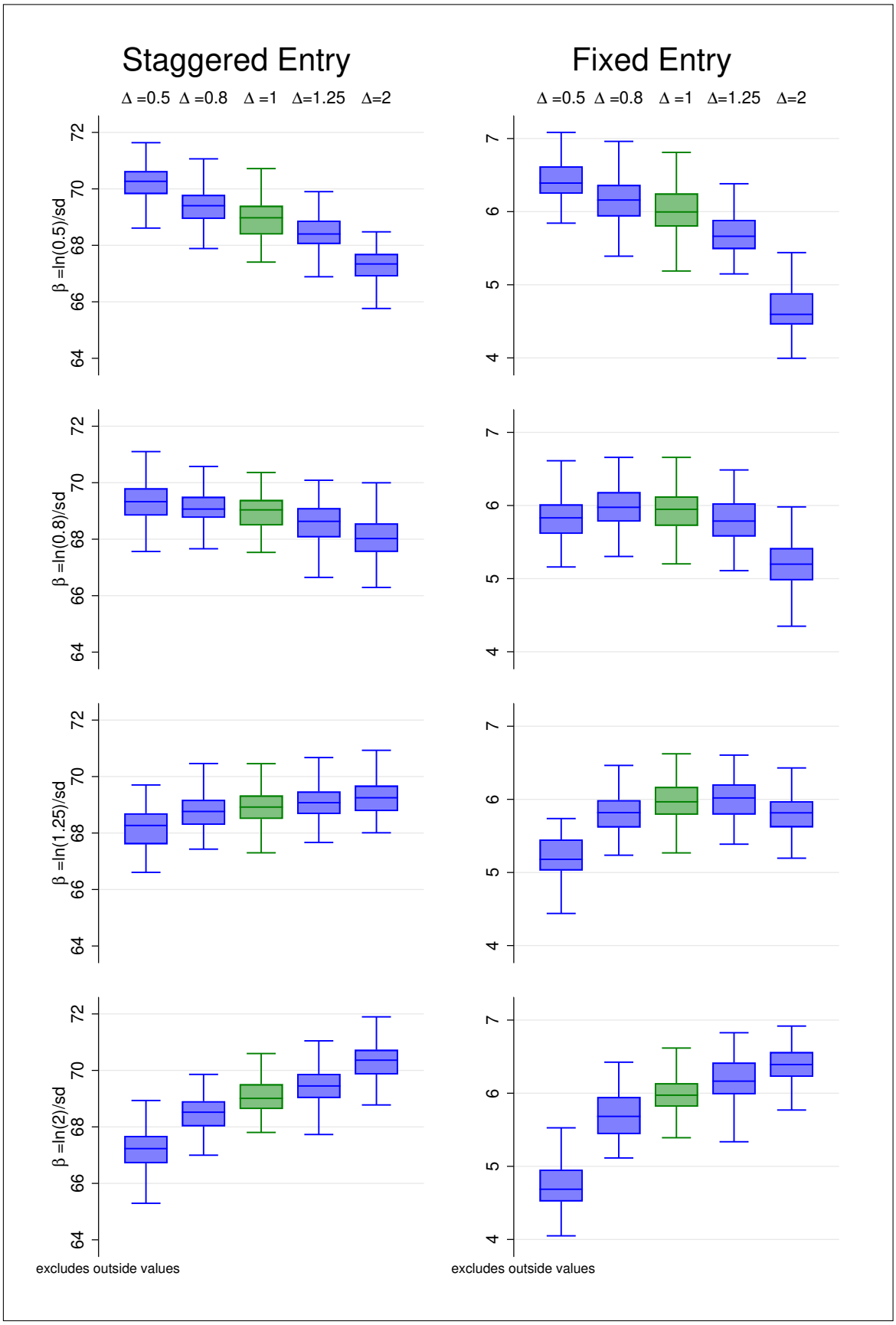


Δ change in hazard ratio between 25th and 75th percentile of reference survival times



Δ change in hazard ratio between 25th and 75th percentile of reference survival times

Figure 2.4: Median Survival Times of Cases for varying Δ and, varying β in the reference dataset



Δ change in hazard ratio between 25th and 75th percentile of reference survival times

2.8 Discussion

The methods described in this chapter bring in their own possible confounding factors and limitations.

The choice of a Weibull model for simulation of survival times limits the scope of the results. It is possible that the methods explored in this thesis may exhibit differing behaviours under exponential, Gompertz, or other distributions. Further, the choice of a Weibull model for simulation of survival times requires that under non-proportional hazards the covariate interacts with ln(time). It is possible that interactions with other functions of time may give different results.

The choices to model parameters for the Weibull distribution from the InterAct dataset, to scale λ, and to choose values for ϕ_k and β_{ϕ_k} such that non-proportional hazards datasets show some similar characteristics to the proportional hazards datasets simulated in this thesis, all introduce the concern that the scope of simulation studies in this thesis is limited to datasets that are similar to InterAct. However, in any simulation study design, limiting choices must be made, and the advantages of consistency across simulation studies, and applicability to at least one example of a real-world case-cohort dataset motivated these decisions.

There are alternatives for the censoring and sampling methods described in this chapter. Administrative censoring could be carried out at a specified follow-up time, based on some model that is expected to give the desired case percentage. Subcohort sampling could be carried out by selecting as the subcohort all subjects with $U \leq \alpha$, as the CDF of a $U(a, b)$ distribution is $\frac{x-a}{b-a}$. Both such alternatives would introduce a level of variation in the case percentages and sampling fractions seen in each simulated dataset. As part of the aim of this data-generating mechanism is to minimise any potential confounding effects, it was decided not to introduce such variation in this thesis.

In practice, recruitment and, in some cases, administrative censoring, will likely take place over a period of time, for logistical reasons, whereas in this data-generating mechanism, recruitment and administrative censoring are instantaneous. The fixed entry data generating mechanism is therefore a "best case scenario", with no loss of subjects due to staggered entry, staggered exit or loss-to-follow-up. Note, however, that for evaluating whether estimation and post-estimation methods are appropriate in the case-cohort design, such a "best case scenario" may be valuable.

Chapter 3

Estimation

3.1 Introduction

In the case-cohort design, estimation under the Cox PH model requires adjustments for the over-representation of cases in the case-cohort sample. Going forward, such adjustments will be referred to as weighting methods.

A number of weighting methods, as described in Chapter 1, have been proposed. In section 3.2, I describe the weighting methods for estimation of β considered in this chapter. As described in Chapter 1, in practice, the most commonly used weighting methods in the case-cohort design are those proposed by Prentice (1986) and Barlow (1994) (Sharp et al., 2014). Prentice and Barlow weighting are hence considered due to their wide use in practice. Inverse probability weighting (IPW), is also considered due to its ease of implementation and reports from literature that it may have improved performance over Prentice and Barlow in certain circumstances. Further, in future chapters, IPW may be more easily applied where weighting is required for post-estimation procedures to be adapted to the case-cohort design. Application of time-based post-stratification to weighting methods is also described in this section.

A number of simulation studies comparing the performance of weighting methods for estimation of β exist in the literature. Kim (2014) attempts to reconcile these conflicting results with the explanation that the relative performance of weighting systems depends upon the interplay between full cohort size, full cohort case-percentage, and subcohort sampling fraction. However, I hypothesise that the relationship is somewhat more nuanced. In Section 3.2.1.5, I briefly describe these comparisons and theorise that relative performance of case-cohort weighting systems depends upon the interplay of censoring type, subcohort size, sampling fraction, and case to non-case ratio in the case-cohort sample.

In Section 3.3, I describe the weighting methods for estimation of cumulative baseline hazard $H_0(t)$ considered in this chapter. In section 3.4, I perform a simulation study comparing the performance of these weighting methods in a variety of circumstances. In section 3.6, I discuss the results of the simulation study, reconcile my results with the extant literature, and make recommendations for estimation of β and $H_0(t)$ in the case-cohort design.

Overall, results from the simulation study indicated that both IPW and Prentice are appropriate for estimation of β in the case-cohort design, and offer improved performance over Barlow. IPW weighting shows similar performance to Prentice in estimation of $H_0(t)$, with post-stratification based on time providing an improvement in precision for both weighting methods, particularly where the sampling fractions of non-cases in the risk sets for each failure time show high variability.

3.2 Estimation of Coefficients

3.2.1 Weighting Methods

Recall from Chapter 1, that, using the notation and risk set structure from 1.2.1.4, the partial likelihood for the Cox model is given by

$$pL(\beta) = \prod_{j=1}^{N^C} \frac{exp(\beta^T Z_{[j]})}{exp(\beta^T Z_{[j]}) + \sum_{i \in R_{(j)}} exp(\beta^T Z_i)}$$

and the pseudopartial likelihood for a case-cohort sample is given by

$$pL^*(\beta) = \prod_{j=1}^{N^C} \frac{w_{[j]} exp(\beta^T Z_{[j]})}{w_{[j]} exp(\beta^T Z_{[j]}) + \sum_{k=1}^{K} w_{(j)}^k \sum_{i \in R_{(j)}^k} exp(\beta^T Z_i)}$$

Let the subcohort sample refer to a randomly selected sample of the full cohort, with each observation having probability of selection p. Let the case-cohort sample refer to the subcohort sample and any full-cohort cases not included in the subcohort sample. Let $\alpha_{(j)}$ refer to the subcohort sampling fraction for the risk set associated with each failure time. i.e. the proportion of the full cohort risk set $R_{(j)}$ included in the subcohort sample. Let α, the subcohort sampling fraction at time $t_{(0)}$, refer to the proportion of the full cohort included in the overall subcohort sample. Let $R_{(j)}$ $= R_{(j)}^C \cup R_{(j)}^{NC}$, the union of the cases at risk at time $t_{(j)}$ and the non-cases at risk at time $t_{(j)}$. Hence, let α^C and α^{NC}, refer to the subcohort *non-case* and subcohort

case sampling fractions, and let $\alpha_{(j)}^{C}$ and $\alpha_{(j)}^{NC}$ refer to the proportion of the full cohort cases and non-cases at risk at time $t_{(j)}$ at risk in the subcohort at time $t_{(j)}$, respectively.

3.2.1.1 Prentice Weighting

Analysis of the case-cohort sample in the Cox PH model, as proposed by Prentice (1986), is carried out via a modification of the partial likelihood. Non-subcohort cases are excluded from the denominator risk sets at all times except for at their time of failure. This can be accomplished by giving subcohort observations weight 1 at all times, with non-subcohort cases taking weight 0 prior to failure, and weight of 1 at failure time $t_{(j)}$. Alternatively, entry time of non-subcohort cases can be altered so that the non-subcohort cases enter the study just prior to their failure time.

3.2.1.2 Barlow Weighting

Barlow weighting treats non-subcohort cases as in Prentice weighting, with a weight of 0 prior to failure, and a weight of 1 at their failure time $t_{(j)}$. Subcohort cases are also given weight 1 at their failure time $t_{(j)}$, and all other subcohort observations at risk at time $t_{(j)}$ are weighted by the inverse of the subcohort sampling fraction. Two separate variations for the subcohort sampling fraction have been proposed. Originally, Barlow (1994) proposed use of the subcohort sampling fraction at time $t_{(j)}$, $\alpha_{(j)} = \frac{N_{(j)}^{SC}}{N_{(j)}}$. However, as this required different weights at each failure time, with the subcohort being enumerated each time, Barlow et al. (1999) suggests use of the overall subcohort sampling fraction, $\alpha = \frac{N^{SC}}{N}$ as an estimator for $\alpha_{(j)}$. This appears to be common practice, however given advances in computing, this repeated enumeration is no longer such a barrier should it result in improved performance.

For ease of description, let the Barlow weighting method where α is used for weighting of $i \in R_j^*$ be termed Barlow Classic and let the Barlow weighting method where $\alpha_{(j)}$ is used for weighting of $i \in R_j^*$ be termed Barlow Time.

3.2.1.3 Inverse Probability Weighting

Inverse probability weighting (IPW) gives each observation a weight inverse to its probability of inclusion in the analysis. In the case-cohort design, while both cases and non-cases are included in the *subcohort sample* with probability p, cases are included in the *case-cohort sample* with probability 1. Hence, use of a weight of 1 for all cases in the case-cohort sample can be justified under IPW analysis. The risk set for *cases* is the same as that of the full cohort and variation occurs only with regard to the selection of subcohort non-cases.

This was first explicitly proposed for case-cohort analysis by Kalbfleisch and Lawless (1988) with $p = \alpha$ for non-cases. Chen and Lo (1999) also proposed a case-cohort estimator which can be considered to be an IPW variant with $p = \alpha^{NC}$ for non-cases, however it is not presented as a form of IPW. For the stratified case-cohort design, Borgan et al. (2000) describes a number of IPW weighting methods, including that of Chen extended to the stratified case-cohort design, and an adaptation where non-cases are the above weights are replaced by their time-dependent equivalents $\alpha_{(j)}^{NC}$.

For ease of description, let the IPW weighting method where α^{NC} is used for weighting of $i \in R_j^{*NC}$ be termed IPW Classic and let the IPW weighting method where α_j^{NC} is used for weighting of $i \in R_j^{*NC}$ be termed IPW Time.

3.2.1.4 Post-Stratification Approach

Both Barlow and IPW weighting methods can be considered as an attempt to mimic or duplicate the denominator of the full cohort likelihood, with IPW using the additional information of the cases outside the subcohort at times other than their failure time. Note that this is not a requirement for correct estimation, but is a similarity between the methods. Unlike Prentice weighting, under Barlow and IPW weighting the expectation for the denominator of the pseudolikelihood at each failure time $t_{(j)}$ is equal to that of the full cohort, since the subcohort case and non-case risk sets at each failure time $t_{(j)}$ can be considered as a simple random sample from the full cohort cases or non-cases at that failure time $t_{(j)}$.

For a *particular* failure time $t_{(j)}$, IPW Classic and Barlow Classic are likely to assign weights that do not reflect the true composition of the full cohort at that time. This is due to variation in the empirical overall subcohort sampling fractions and non-case subcohort sampling fractions, as their sizes at particular times $t_{(j)}$ vary. Essentially, the use of α and α^{NC} to estimate $\alpha_{(j)}$ and $\alpha_{(j)}^{NC}$, respectively, likely means that there are elements of the variation in the denominator of the pseudolikelihood which can be compensated for by adjusting weights based on empirical sampling fractions at each failure time $t_{(j)}$. While this does not directly mean there will be a decrease in variation of $\hat{\beta}$, a reduction in variance of this denominator might imply an improvement in precision of $\hat{\beta}$.

Barlow Time and IPW Time can be considered as post-stratification approaches that make use of available information on substrata of the full cohort so as to reduce variation in the denominator of the pseudolikelihood, accounting for the variation in the ratio of subcohort size to full cohort size, and subcohort non-cases to full cohort

non-cases at a particular time $t_{(j)}$, respectively. This approach can be generalised to any definition of potentially relevant strata where the number of observations within each stratum is known for both the full cohort and the subcohort sample.

Table 3.1 summarizes variant weights for each component of the pseudolikelihood.

Table 3.1: Weights for Each Component of the Pseudolikelihood at risk at time t(j)

	Prentice	Barlow Classic	Barlow Time	IPW Classic	IPW Time
Case failing at time $t_{(j)}$	1	1	1	1	1
Non-Subcohort Case at risk at time $t_{(j)}$	0	0	0	1	1
Subcohort Case at risk at time $t_{(j)}$	1	$1/\alpha$	$1/\alpha_{(j)}$	1	1
Subcohort Non-Case at risk at time $t_{(j)}$	1	$1/\alpha$	$1/\alpha_{(j)}$	$1/\alpha^{NC}$	$1/\alpha_{(j)}^{NC}$

α overall subcohort sampling fraction; α^{NC} overall subcohort non-case sampling fraction; $\alpha_{(j)}$ subcohort sampling fraction at $t_{(j)}$; $\alpha_{(j)}^{NC}$ subcohort non-case sampling fraction at $t_{(j)}$

3.2.1.5 Comparative Performance of Weighting Methods

Some studies exist that compare the relative performance of Prentice, Barlow, and/or IPW methods. Drawing overarching conclusions from these studies is difficult, as each study considers different combinations of full cohort sizes, case percentages, sampling fractions, and analysis time scales. Reporting of results as they relate to various case-cohort characteristics of interest is also not consistent. Datasets are generally described in terms of full cohort size, subcohort sampling fraction, and case percentage, which allows for understanding of performance of weighting systems as compared to the full cohort. However, for comparison of performance between weighting systems, and guidance to end-users as to which weighting method to use, subcohort size (N^{SC}) and non-case to case ratio in the case-cohort sample may be more relevant and estimates of these quantities are also presented here.

I theorise that relative performance of case-cohort weighting systems depends upon the interplay of entry and exit type, subcohort size, sampling fraction, and case to non-case ratio. Under fixed entry analysis, with no loss to follow-up, the maximum subcohort risk set size is the size of the subcohort, with risk set size decreasing with time, as subcohort cases reach their failure time. The minimum subcohort risk set size is the size of the subcohort non-cases. With loss-to-follow-up, the maximum subcohort risk set size is the same, and risk set size still decreases with time, but to a greater degree, as a portion of the subcohort non-cases are absent from risk sets after reaching their censoring time. Under staggered entry, subcohort risk set sizes will be smaller and more variable than if the same dataset was analysed under fixed

entry. Size of subcohort risk sets will decline steadily over time, as subjects display delayed entry. As subcohort size and sampling fraction decrease, subcohort risk set sizes become more variable. As non-case to case ratio increases, Barlow and Prentice weighting lose more information from exclusion of non-subcohort cases from risk sets.

In this section, the characteristics of each of the studies are first summarized, followed by an overall summary of their results and conclusions.

3.2.1.5.1 Characteristics of Studies Comparing Weighting Methods

Barlow et al. (1999) compares Barlow Classic and Prentice in the Welch Nickel Refinery dataset, which is a full cohort of 637 subjects with $\sim 9\%$ cases, analysed with age as time-scale (staggered entry). Sampling fractions from 10%-90% were considered, each drawn from the the Welch Nickel Refinery dataset over 200 replicates. N^{SC} was therefore ~ 64 to 570, with non-case to case ratios from 1:1 to 9.2:1.

In a simulation study comparing estimators in the stratified case-cohort design, Borgan et al. (2000) also compare Prentice and IPW Classic in an unstratified sample from a full cohort of 1000 subjects with 10% cases and 10% sampling fraction over 1000 replicates, analysed with fixed entry and 20% early censoring. N^{SC} was therefore ~ 100 with non-case to case ratio 1:1.

Petersen et al. (2003) compares Barlow Classic, Prentice, and IPW Classic in full cohorts of 12,301 with 13.6% cases, analysed with age as time-scale (staggered entry). Sampling fractions of 3%, 11%, and 20% percent were considered, each over 1000 replicates. N^{SC} was therefore $\sim 370, 1350, 2500$, with non-case to case ratios 0.2:1, 0.7:1 and 1.3:1, respectively.

Onland-Moret et al. (2007) compares Barlow Classic and Prentice under a number of scenarios, each analysed with fixed entry over 50 replicates. Type of right censoring was not explicitly specified, but appears to be fixed administrative censoring. Scenarios considered were 8% cases and full cohorts of 1,000 for sampling fractions ranging from 5%-70%; 8% cases and sampling fractions of 5%, 10% and 20% in full cohorts ranging from 200-2000; sampling fraction of 10% and full cohort of 2000 with case percentages ranging from 1-30%. N^{SC} was therefore $\sim 10, 20, 40, 50, 100$, with non-case to case ratio 0.6:1, 1.2:1, 2.3:1, 0.6:1, and 0.6:1, respectively; N^{SC} ~ 200 with non-case to case ratios 0.2:1, 1.2:1, and 9.9:1; $N^{SC} \sim 400$ with non-case to case ratio 2.3:1; and $N^{SC} \sim 700$ with non-case to case ratio 8.1:1.

Kim (2014) compares Barlow Classic, Prentice, and IPW Classic under a number of

scenarios, each analysed with fixed entry over 5000 replicates. Right censoring was carried out by generating random censoring times, uniformly distributed between 0 and c, with c chosen such that the desired case percentage would be achieved. The earlier of censoring or failure time was observed for each subject. This study hence included loss-to-follow-up, though the degree of this was not specified. Scenarios considered were 15% cases in full cohorts of 500, 1000, and 1500, and sampling fractions of 14%, 25% and 49%; and case percentages p_C of 0.1%, 1%, 5%, 10%, each with sampling fraction 14%, 25%, 49% and full cohorts equal to $7500p_C$, $15000p_C$, and $22500p_C$. The simulation study with 15% cases was also conducted with fixed administrative censoring times.N^{SC} therefore ranges from $\sim 70-200$ with non-case to case ratio 0.8:1 to 2.8:1; N^{SC} from $\sim 250-750$ with non-case to case ratio 1.3:1 to 9.3:1; N^{SC} from $\sim 1000-7500$ with non-case to case ratio 4.4:1 to 48.5:1; and N^{SC} from $\sim 10000-110000$ with non-case to case ratio 48.5:1 to 489.5:1.

3.2.1.5.2 Results for Barlow vs. Prentice

In no comparison does Barlow offer improved performance over Prentice. Petersen found no differences in bias or efficiency between Barlow Classic and Prentice Classic. Onland Moret suggests that where the subcohort is 15% or more of the full cohort size, Barlow Classic and Prentice give extremely similar results, that are also very similar to the full cohort estimates. A number of comparisons found that Barlow overestimates the coefficients where subcohort sample sizes are small (between 90 and 140 subjects, depending on sampling fraction and full cohort size). Further, a number of comparisons found that Barlow showed higher variance of coefficient estimates when subcohort size was small (100 subjects in Onland Moret, 200 subjects in Barlow, and 140 subjects in Kim. It should be noted that Kim did not specify the relative performance of Prentice and Barlow in the study with fixed administrative censoring, and this finding is for the study with random right censoring. In general, where differences were found, it appears that the magnitude of the hazard ratios, and use of binary or continuous covariates makes little difference to the relative performance of Barlow Classic and Prentice weighting.

3.2.1.5.3 Results for IPW vs. Barlow/Prentice

Petersen found no differences in bias or efficiency between Prentice, and IPW Classic. Borgan found that IPW Classic displays improved efficiency over Prentice where the covariate is more variable or has a greater mean. Kim found that IPW Classic yielded higher efficiency than Prentice and Barlow, with this increase in efficiency greater at higher magnitude of β. Kim also found that IPW Classic had greater power than Prentice, but the relative difference of power was moderate even where difference of variance was relatively large. Kim also found that as the proportion of

failure events in the full cohort became smaller ($< 10\%$), differences in performance between weighting methods reduced. Overall, the literature suggests that IPW Classic weighting may show similar or improved performance compared to Prentice weighting, and improved performance compared to Barlow Weighting.

3.3 Estimation of Cumulative Baseline Hazard

Recall from Chapter 1 that, using the notation and risk set structure from Chapter 1 Section 2.1.4, the the Breslow estimator of $H_0(t)$ in the full cohort is given by:

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \hat{h}_0(t_{(j)}) = \sum_{t_{(j)} \leq t} \frac{1}{exp(\beta^T Z_{[j]}) + \sum\limits_{i \in R_{(j)}} exp(\beta^T Z_i)}$$

and the weighted Breslow estimator of $H_0(t)$ in the case-cohort is given by:

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \hat{h}_0(t_{(j)}) = \sum_{t_{(j)} \leq t} \frac{1}{w_{[j]} exp(\beta^T Z_{[j]}) + \sum\limits_{k=1}^{K} w^k_{(j)} \sum\limits_{i \in R^k_{(j)}} exp(\beta^T Z_i)}$$

Prentice (1986) proposes a case-cohort estimator of cumulative baseline hazard where non-subcohort cases are excluded from the denominator risk sets at all times except for at their time of failure. The case failing at time $t_{(j)}$ and the subcohort observations at risk are weighted by the inverse of the subcohort sampling fraction α. Kulich and Lin (2004) propose a general class of weighted estimators for $H_0(t)$, of which IPW Classic and IPW Time can be considered special cases.

In this Chapter, I consider IPW Classic, IPW Time, Prentice's estimator (Prentice Classic), and an adaptation Prentice Time, extending the post-stratification based on the ratio of subcohort size to full cohort size at a particular time $t_{(j)}$ to Prentice's estimator. In Prentice Time, the case failing at time $t_{(j)}$ and the subcohort observations at risk are weighted by the inverse of the subcohort sampling fraction at that time $t_{(j)}$, $\alpha_{(j)}$.

Sanderson et al. (2013) compare Prentice Classic and Barlow Classic estimates of $H_0(t)$ in the context of risk prediction measures and find that both methods provide similar estimates, but that Prentice Classic is slightly more accurate than Barlow Classic. To my knowledge no simulation study comparing the performance of IPW and Prentice estimators of $H_0(t)$ are extant in the literature.

3.4 Simulation Study

This simulation study compares the performance of selected weighting methods to each-other and the full cohort in estimation of coefficients and cumulative baseline hazard in the case-cohort design, in circumstances similar to those that might be seen in practice.

3.4.1 Data Generating Mechanism

Data generation was carried out as per Chapter 2, with the following specifications: For comparison of estimation of coefficients, full cohorts were simulated with three binary covariates drawn independently from $Binomial(1, 0.5)$ and three continuous covariates drawn independently from $Normal(0, 1)$, for six covariates per dataset. Survival times were generated with vector of coefficients for each covariate type equal to hazard ratios (HR) per standard deviation (SD) of 1, 1.25, and 2, respectively. For comparison of estimation of cumulative baseline hazard, full cohorts were simulated with a binary covariate drawn independently from Binomial(1,0.5) and a continuous covariate drawn independently from Normal(0,1), for a total of two covariates per dataset. Coefficients for each covariate are equal, with two coefficients considered, equal to hazard ratios per standard deviation of 1.25 and 3.

3.5 Estimands

For estimation of coefficients, a Cox PH model was fitted using each weighting method of interest, with all six covariates included in the model. Where weights varied by time, weights were rounded to one significant digit to reduce computation time. Coefficient estimates and their associated standard errors were recorded for each weighting system and the full-cohort. For estimation of $H_0(t)$, a Cox PH model was fitted using Prentice Classic weighting, with both covariates included in the model. Case-cohort estimates of $H_0(t)$ were calculated using the Prentice Classic estimate of β and the Breslow estimator weighted appropriately for the weighting system. For fixed entry, $H_0(t)$ was calculated from time $t_{(0)} = 0$. For staggered entry, $H_0(t)$ was calculated from times chosen so as to avoid many missing estimates early in time, based on the number of cases N^C in the case-cohort dataset.

$$t_{(0)} = \begin{cases} 45, & \text{if } 500 \leq N^C \\ 50, & \text{if } 150 \leq N^C < 500 \\ 55, & \text{if } 75 \leq N^C < 150 \\ 60, & \text{if } N^C < 75 \end{cases}$$

For each combination of time-scale, subcohort size, subcohort sampling fraction, non-case to case ratio, and vector of coefficients, a reference full cohort dataset was generated using the same data-generating mechanism, with the exception that the full cohort was of size 50000. 100 benchmark survival times were recorded for the reference dataset as 100 quantiles of the survival times of cases that survived beyond the appropriate time $t_{(0)}$ above. Estimates of $H_0(t)$ in the case-cohort dataset for the benchmark survival times were carried forward from the previous survival time. Full cohort estimates were also calculated.

3.5.1 Methods

For estimation of coefficients, Prentice, Barlow Classic, Barlow Time, IPW Classic and IPW Time were considered, as well as the full cohort. For estimation of cumulative baseline hazard, Prentice Classic, Prentice Time, IPW Classic and IPW Time were considered, as well as the full cohort.

3.5.2 Performance Measures

Statistics for performance measures were calculated using the `simsum` Stata package (White, 2010), together with their associated Monte Carlo standard errors (MCSE) which quantify the uncertainty due to a finite number of simulations.

For estimation of coefficients performance measures calculated were bias, mean squared error, empirical standard error, power, coverage, and proportional error in the model-based standard error. For estimation of $H_0(t)$, performance measures bias, empirical standard error, and mean square error were calculated at each benchmark survival time.

MCSE bounds for each performance measure statistic were calculated as the statistic +/-1.96*MCSE. Where the MCSE bounds of the weighting methods do not overlap, it indicates a statistically significant difference in the performance of the weighting methods. While this is a conservative assessment, and a lack of statistical significance should not be inferred from overlapping of MCSE bounds, it does provide a level of objectivity.

3.5.3 Results

3.5.3.1 Estimation of Coefficients

Time weighting made only minimal difference and hence these variants are not presented here. All case-cohort estimators performed similarly for Power, Type

1 error, coverage, and proportional error in the model-based standard error. In general, Type 1 error was close to a nominal 5%, coverage was close to a nominal 95%, and proportional error in the model-based standard error was within 5% of 0. Where this pattern was not followed, it was usually in combinations of the smaller subcohort size of 200, smaller 3% sampling fraction, and/or staggered entry. Further details on these results are included in the Appendix. Results for Bias, empirical standard error (ESE) and mean square error (MSE) are discussed below. Graphs of these results for the subcohort of size 200 presented in Figure 3.1.

3.5.3.1.1 Bias Under the null, MCSE bounds for bias of all weighting systems overlap with the full cohort and eachother, and generally encompass 0. At subcohort size 1000, point estimates for bias do not exceed $\sim$ +/-4% of the true β, upper MCSE bounds do not exceed 6.5% of the true β and lower MCSE bounds do not exceed $\sim$ +/-4% of the true β. As β increases, case-cohort estimators tend to display more bias away from the null, especially at the 3% sampling fraction and subcohort size 200. This is not unexpected, as in small samples $\hat{\beta}$ has a heavy-tailed distribution. At subcohort size 200, point estimates do not exceed 17%, 16% and 10% of the true β, for Barlow, IPW, and Prentice, respectively. Their respective upper MCSE bounds do not exceed 21%, 19%, and 15%, and lower MCSE bounds do not fall below 1%, 0%, and -4%. Barlow generally displays greatest bias, and Prentice the smallest. Differences in bias between Prentice and IPW exceed MSE bounds for the Binary covariate with $\beta = \ln(2)/0.5$, sampling fraction 3%, and staggered entry.

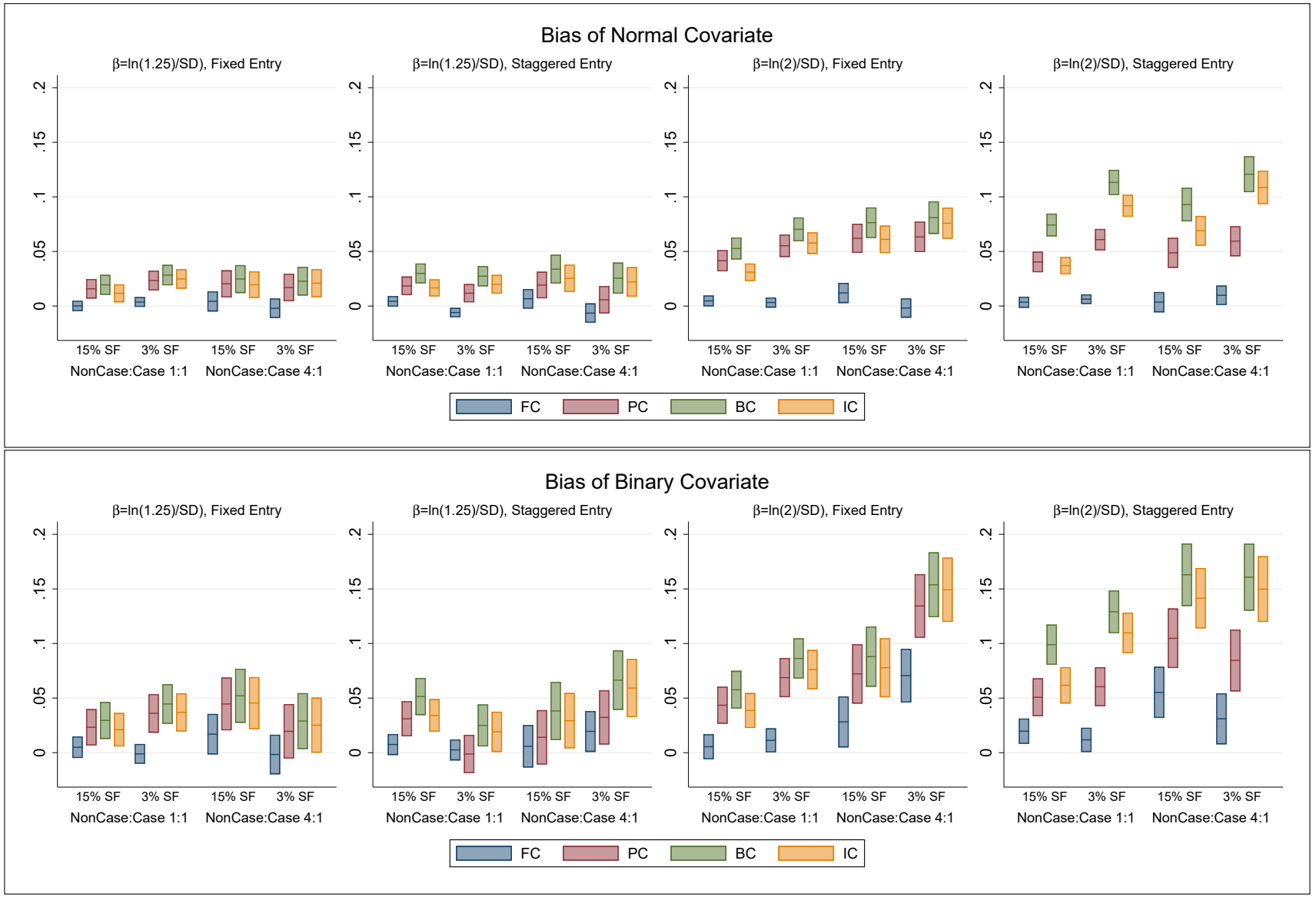
3.5.3.1.2 Empirical Standard Errors MCSE bounds for ESE of case-cohort weighting systems tend to overlap, although Barlow tends to show higher ESE than IPW and Prentice. IPW tends to show lower ESE than Prentice at the 15% sampling fraction, and Prentice tends to show lower ESE than IPW at the 3% sampling fraction and subcohort size 200, though these differences do not exceed MCSE bounds.

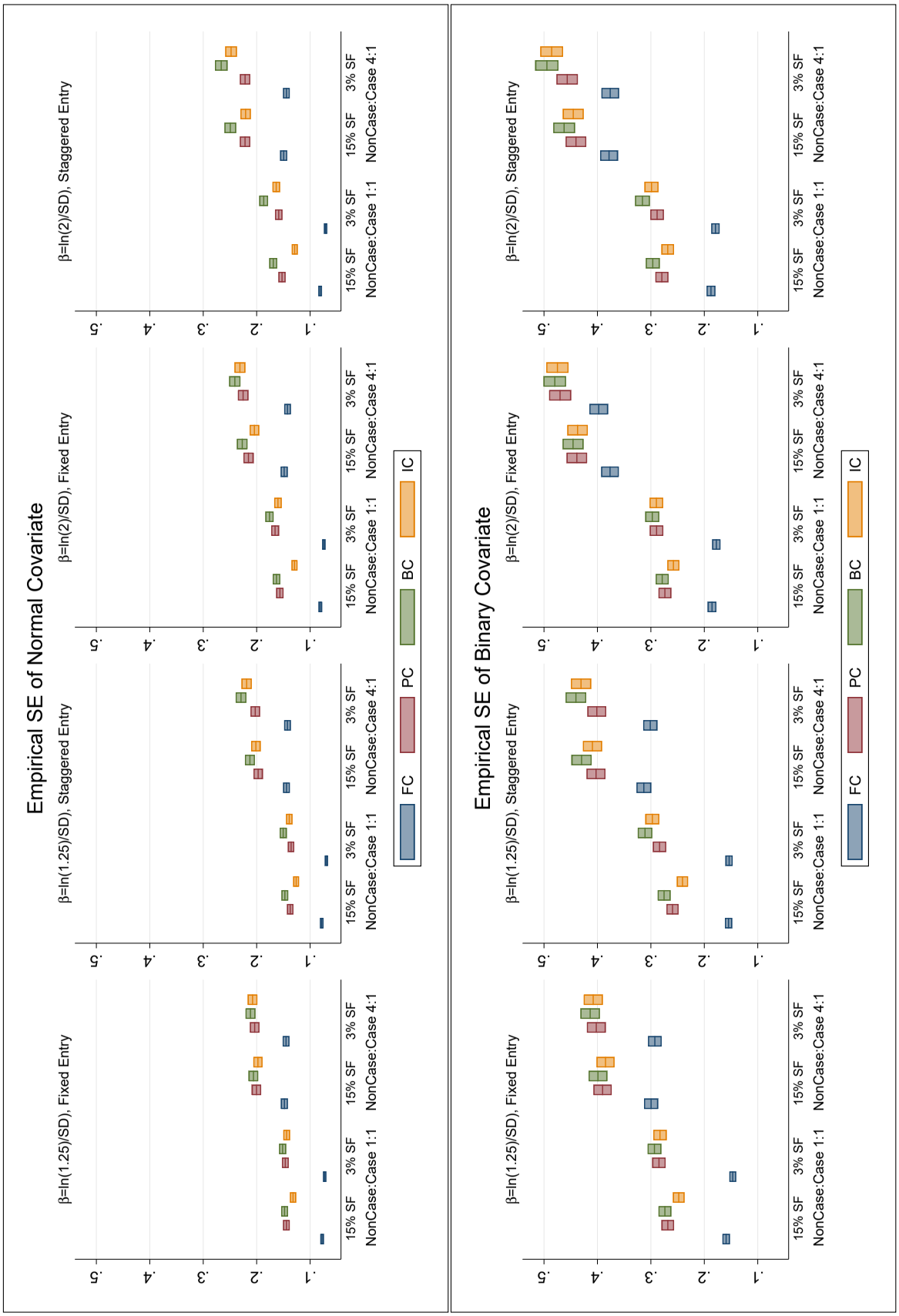
3.5.3.1.3 Mean Square Error MSE reflect the previous results; Barlow tends to show highest MSE, Prentice the lowest at the 3% sampling fraction and subcohort size 200 and IPW the lowest at the 15% sampling fraction. The differences in MSE between Prentice and IPW cause MCSE bounds to fail to overlap for the Normal covariate with $\beta = \ln(2)$, sampling fraction 15% and case to non-case ratio 1:1.

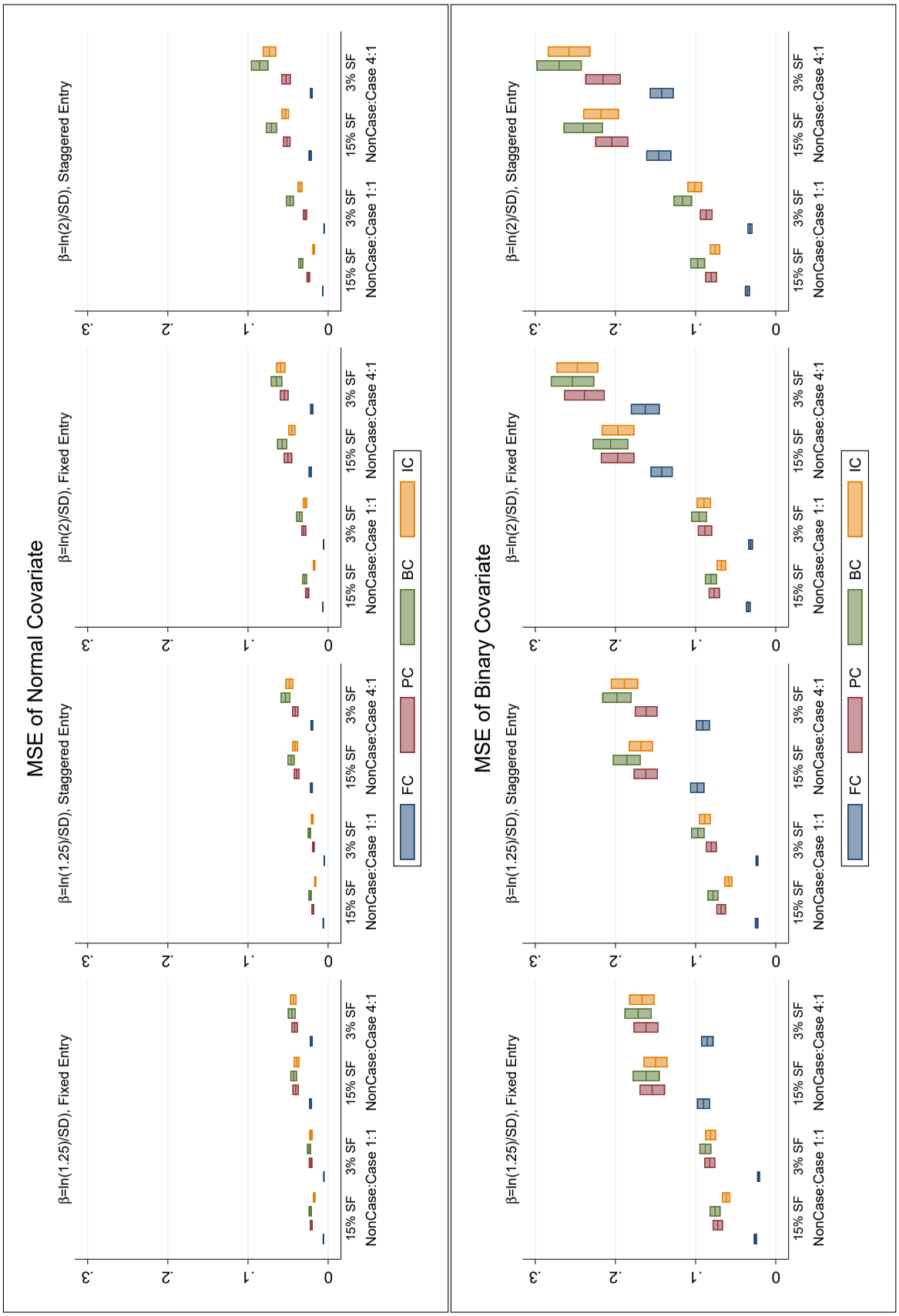
Figure 3.1: Bias, Empirical Standard Error, and Mean Square Error of Estimates of β

FC= Full Cohort; BC = Barlow Classic; PC = Prentice Classic; IC = IPW Classic

Shaded bars indicate 95% CI reflecting Monte Carlo error




Empirical SE of Normal Covariate
β=ln(1.25)/SD), Fixed Entry
β=ln(1.25)/SD), Staggered Entry
β=ln(2)/SD), Fixed Entry
β=ln(2)/SD), Staggered Entry
.1
.2
.3
.4
.5
15% SF
3% SF
15% SF
3% SF
NonCase:Case 1:1
NonCase:Case 4:1
FC
PC
BC
IC
Empirical SE of Binary Covariate
β=ln(1.25)/SD), Fixed Entry
β=ln(1.25)/SD), Staggered Entry
β=ln(2)/SD), Fixed Entry
β=ln(2)/SD), Staggered Entry
.1
.2
.3
.4
.5
15% SF
3% SF
15% SF
3% SF
NonCase:Case 1:1
NonCase:Case 4:1
FC
PC
BC
IC


MSE of Normal Covariate
β=ln(1.25)/SD), Fixed Entry
β=ln(1.25)/SD), Staggered Entry
β=ln(2)/SD), Fixed Entry
β=ln(2)/SD), Staggered Entry
MSE of Binary Covariate
β=ln(1.25)/SD), Fixed Entry
β=ln(1.25)/SD), Staggered Entry
β=ln(2)/SD), Fixed Entry
β=ln(2)/SD), Staggered Entry
0
.1
.2
.3
15% SF
3% SF
15% SF
3% SF
NonCase:Case 1:1
NonCase:Case 4:1
FC
PC
BC
IC

3.5.3.2 Estimation of $H_0(t)$

Minimum and maximum true values for $H_0(t)$ in each scenario are included in the Appendix.

3.5.3.2.1 Fixed Entry

In fixed entry, post-stratification based on time does not apply to IPW, and for Prentice resulted in performance measures that were almost identical to Prentice Classic, except for a small increase in bias in the latter half of reference times, the difference not exceeding 0.5% of the true value of $H_0(t)$. IPW Classic showed greater bias than Prentice Classic, underestimating the true hazard, and greater empirical standard error. These differences did not exceed 1.2% of the true value of $H_0(t)$ for bias and 2.5% of the true value of $H_0(t)$ for empirical standard error, and did not cause MCSE bounds of the performance measures to fail to overlap.

Overall, in fixed entry, all case-cohort estimators perform similarly to the full cohort, particularly at the higher subcohort size. Where sampling fraction and subcohort size are small, and β and case to non-case ratio is high, case-cohort estimators underestimate the true hazard, with MCSE bounds of bias falling outside MCSE bounds of the full cohort. Empirical standard error for case-cohort estimators is similar to the full cohort at $\beta = \ln(1.1)/\text{SD}$. At $\beta = \ln(2)/\text{SD}$ and case to non-case ratio 1:1, case-cohort MCSE bounds for empirical standard error fail to overlap with those of the full cohort. Results for MSE reflect those of empirical standard error.

Except for subcohort size 200 and case to non-case ratio 1:4, bias for full cohort and case-cohort estimators is less than 6% of the true $H_0(t)$ at all reference times. At subcohort size 200 and case to non-case ratio 1:4, bias is of larger magnitude towards the beginning and end of the timescales, particularly at the 3% sampling fraction in the last 20 reference times where it reaches 16% of the true $H_0(t)$. However, this bias late in time is similar in the case-cohort and the full cohort, and is likely an artefact of the analysis method. Recall that $H_0(t)$ estimates for unobserved reference times were carried forward from the previous failure time in the replicate.

Empirical standard error is generally not substantial compared to the true $H_0(t)$, except at the beginning of analysis time (first 10 reference times) where few events have occurred. Excluding those first 10 reference times, empirical standard error does not exceed 10%, 25% and 50% of true $H_0(t)$ for all scenarios at subcohort size 1000, subcohort size 200 and case to non-case ratio 1:1, and subcohort size 200 and case to non-case ratio 1:4, respectively.

3.5.3.2.2 Staggered Entry

Figure 3.2 demonstrates the effect of risk sets comprised only of cases (case-only risk-sets) in subcohorts of size 200. With staggered entry, a number of replicates displayed such risk sets early and late in time in the case-cohort samples. Exploratory analysis indicated that the presence of case-only risk sets did not affect performance of weighting methods in estimation of coefficients. However, exploratory analysis indicated that presence of case-only risk sets had a substantial effect on performance of IPW estimators of $H_0(t)$ late in time, and a lesser effect under Prentice weighting variants. In this simulation study, such case-only risk sets occurred only towards the very end of the reference analysis times, For the remainder of the results, reference times 95+ are excluded from analysis under staggered entry.

Figure 3.2: Effect of Case-Only Risk Sets on estimates of $H_0(t)$

PC = Prentice Classic, PT = Prentice Time, IC - IPW Classic, IC = IPW Time

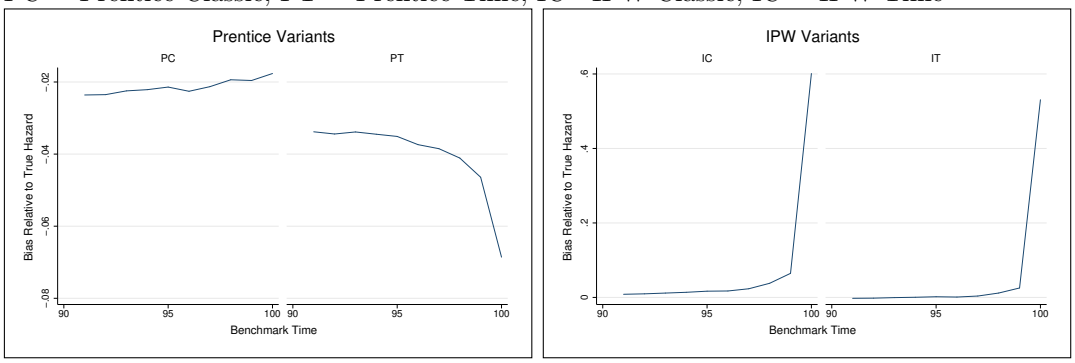


Figure 3.3 shows bias, empirical standard error and mean squared error under staggered entry for all estimators, relative to the true $H_0(t)$ for $\beta = \ln(2)/\text{SD}$, case to non-case ratio 1:4, subcohort size 1000, sampling fraction 15%; and $\beta = \ln(1.1)/\text{SD}$, case to non-case ratio 1:1, subcohort size 200, and sampling fraction 3%.

As in section 3.5.3.2.1, the following discussion of results does not include the beginning of analysis time (first 10 reference times) where few events have occurred. Further detail on the results discussed below can be found in the appendix.

In staggered entry, Prentice Time displays more bias then Prentice Classic, somewhat underestimating the true hazard, particularly towards the end of analysis time. However, this difference in bias does not exceed 1.5% of true $H_0(t)$. Difference in bias between IPW Classic and IPW Time is more variable, ranging from -1.5% of true $H_0(t)$ to +1.4% of true $H_0(t)$.

IPW variants show similar bias to the full cohort, with MCSE bounds overlapping those of the full cohort in all scenarios except at the very end of analysis time. MCSE bounds for bias of Prentice variants, however, fail to overlap with those of the full cohort at the smaller subcohort, with this occuring in more scenarios for Prentice Time than Prentice Classic. At subcohort size 1000, bias for full cohort and all case-cohort estimators is less than 4.2% of the true $H_0(t)$ at all reference times.

At subcohort size 200, bias for the full cohort and IPW variants is less than 3.2%, 4.2% and 5.9% of the true $H_0(t)$ for case to non-case ratio 1:4 and sampling fraction 15%, case to non-case ratio 1:1, case to non-case ratio 1:4 and sampling fraction 3%, respectively. Bias for Prentice variants is less than 5.2%, 8% and 8.3% of the true $H_0(t)$ for case to non-case ratio 1:4 and sampling fraction 15%, case to non-case ratio 1:1, case to non-case ratio 1:4 and sampling fraction 3%, respectively.

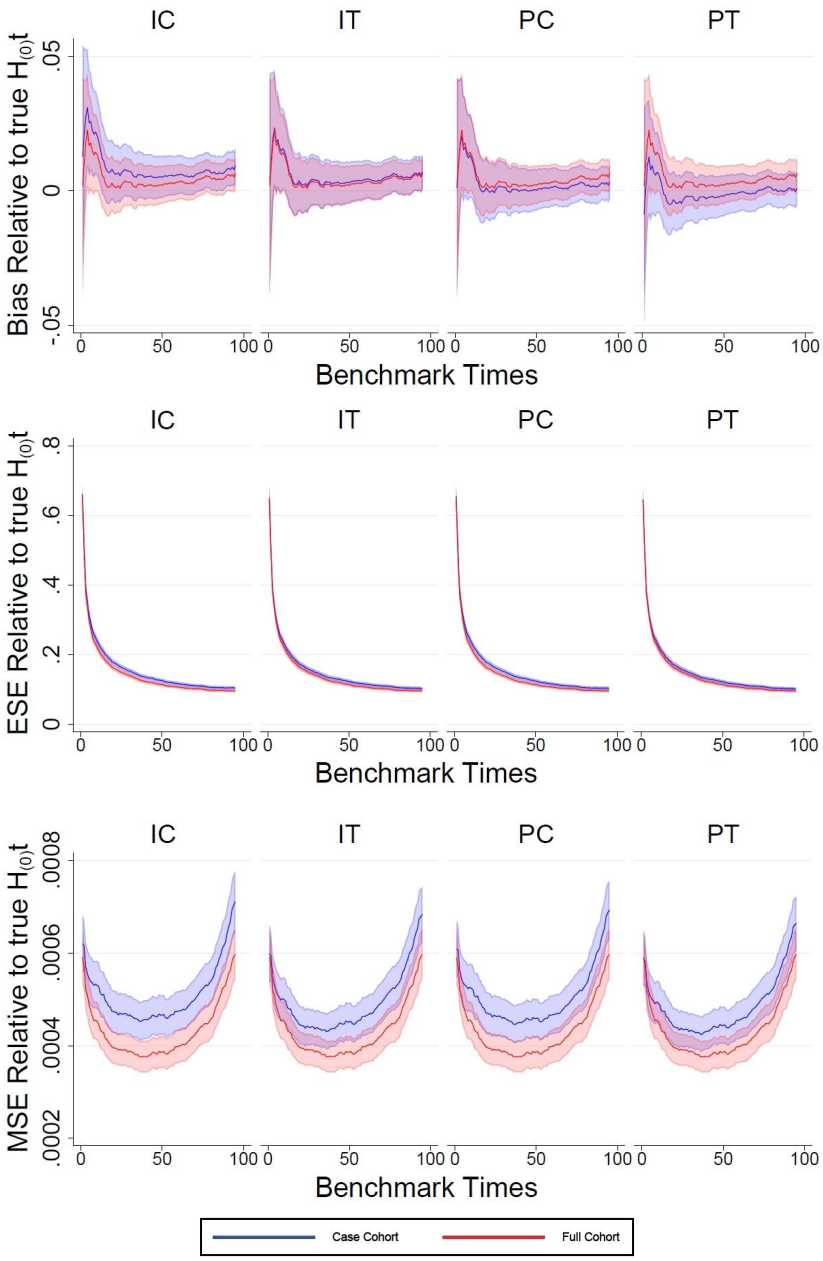
Both Prentice Time and IPW time consistently display lower empirical standard errors than their Classic counterparts, with the difference in empirical standard error higher at smaller sampling fractions, subcohort sizes and β. Difference in empirical standard error between Time and Classic variants does not exceed 3% and 3.9% of true $H_0(t)$ for Prentice and IPW, respectively. This is reflected in MSE with Time variants also displaying lower MSE than their classic equivalents. For both empirical standard error and MSE, MCSE bounds for Classic and Time variants fail to overlap in more than 20% of reference times when $\beta = \ln(1.1)/\text{SD}$, case to non-case ratio is 1:1 and sampling fraction is 3%.

At $\beta = \ln(1.1)/\text{SD}$, IPW Time and Prentice Time show similar empirical standard error to the full cohort, with MCSE bounds overlapping. MCSE bounds for empirical standard error of all case-cohort estimators to overlap with those of the full cohort at $\beta = \ln(2)/\text{SD}$ and case to non-case ratio 1:1. Otherwise, MCSE bounds for empirical standard error of all case-cohort estimators generally overlap with the full-cohort in more than 90% of reference times.

In subcohort size 1000 empirical standard error for full cohort and case-cohort estimators does not exceed 17% and 24% of the true $H_0(t)$ at case to non-case ratios 1:1 and 1:4, respectively. In subcohort size 200 empirical standard error for full cohort and case-cohort estimators does not exceed 31% and 52% of the true $H_0(t)$ at case to non-case ratios 1:1 and 1:4, respectively.

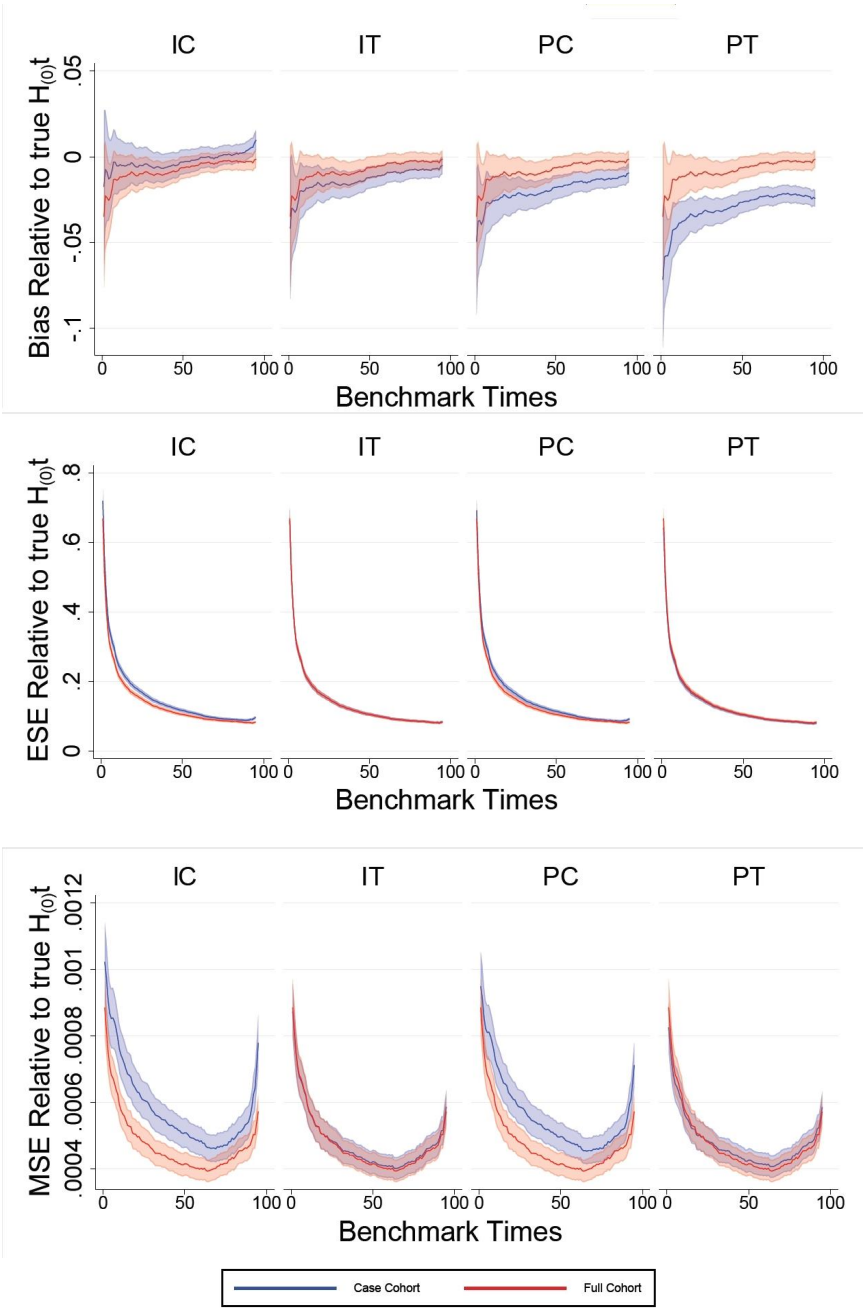
Figure 3.3: Performance Measures for $H_0(t)$

(a) $\beta = \ln(2)/\text{SD}$; non-case to case ratio 4:1; subcohort size 1000; sampling fraction 15%



1

(b) $\beta = \ln(1.1)/\text{SD}$; non-case to case ratio 1:1; subcohort size 200; sampling fraction 3%



3.6 Discussion

3.6.1 Estimation of β

Overall, the simulation study offered no evidence that post-stratification based on time affected the performance of case-cohort estimation of β. While computational speed has advanced since Barlow's paper in 1999, there is still an increase in computational time when enumerating the subcohort at each failure time. Hence, given a lack of evidence to suggest any improvement in performance, post-stratification based on time is not recommended for the case-cohort design. It should be noted that, in the stratified case-cohort design, Borgan reports an improved efficiency for IPW Time over IPW Classic, however, as Borgan notes, this improvement is of little practical relevance, at most 0.2% relative to full cohort efficiency.

The simulation study gives evidence that Prentice shows improved performance over Barlow. Barlow generally displays greater bias, with these differences outside MCSE bounds where β was larger, subcohort size smaller, and under staggered entry. Barlow also showed lower precision than Prentice, with these differences more likely to be outside MCSE bounds where subcohort size, sampling fraction, and non-case to case ratio were smaller. While Petersen found no difference in bias or precision between Barlow and Prentice, the smallest subcohort size considered in that study was 370. Onland Moret suggest that bias of Prentice and Barlow does not differ when sampling fraction reaches 15%, however, this study does not appear to have considered staggered entry. Extant literature shows some disagreement on the subcohort size at which Barlow shows reduced efficiency compared to Prentice, possibly due to differences in data-generating mechanisms between papers. Results from this simulation study are most similar to those of Barlow and Kim. These differing results can be reconciled by considering the interplay of censoring type, subcohort size, sampling fraction, and case to non-case ratio in the case-cohort sample. As already mentioned, Onland Moret did not consider staggered entry. At subcohort size 200, Barlow considered only non-case to case ratio 3.3:1. Kim included loss-to-follow-up, with subcohort size 140 and non-case to case ratio 0.8:1 and subcohort size 250 with non-case to case ratio 1:4.

IPW generally displayed greater bias than Prentice, and less than Barlow, however, the differences in bias between Prentice and IPW caused MCSE bounds to fail to overlap only at sampling fraction 3%, subcohort size 200 and under staggered entry. MCSE bounds of empirical standard error for Prentice and IPW overlapped in all cases, though Prentice tends to show lower empirical standard errors than IPW at the 3% sampling fraction and subcohort size 200, and IPW tends to show lower

empirical standard errors than Prentice at the 15% sampling fraction. The improvement in MSE for IPW over Prentice cause MCSE bounds to fail to overlap for the Normal covariate with $\beta = \ln(2)$, sampling fraction 15% and case to non-case ratio 1:1. In the extant literature, Petersen again found no difference in bias or efficiency between IPW and Prentice, but the non-case to case ratio, sampling fractions, and subcohort sizes considered in that study are discordant to those considered here. Results from this simulations study are similar to those found by Borgan, with the circumstance where IPW shows improved efficiency over Prentice including the normal covariate, which is more variable than the binary covariate. Results from this study at the 15% sampling fraction also reflect the findings of Kim, that IPW yields higher efficiency than Prentice at higher magnitudes of β. However, many of the circumstances considered in Kim, particularly those with very large non-case to case ratios were not considered in this simulation study. While IPW displays a small improvement in power for $\beta = \ln(2)/\text{SD}$ over Prentice at subcohort size 200, sampling fraction 15% and non-case to case ratio 1:1, this improvement does not cause MCSE bounds for these weighting methods to fail to overlap. Kim reports a similar increase in power of 2% for subcohort size 125, sampling fraction 25% and non-case to case ratio 1.4:1.

Overall, results from the simulation study indicated that both IPW and Prentice offer improved performance over Barlow in estimation of β, and that while IPW may show a small increase in bias over Prentice under staggered entry at smaller subcohort sizes and sampling fractions, this does not effect mean squared error. At higher sampling fractions, smaller non-case to case ratios, and with more variable covariates, IPW shows improved efficiency over Prentice. The simulation study indicates that both IPW and Prentice weighting are appropriate for estimation of β in the case-cohort design.

The bias in full cohort estimates is likely due to the small number of observed events, i.e. high censoring proportion of the full cohort, exacerbated where full cohort sizes are small. As described by Kotz Johnson (1985), the Cox estimator is asymptotically unbiased, not unbiased. Persson and Kamis (2005) investigate bias in the Cox model in full cohorts under various circumstances, and find that even under proportional hazards, the Cox estimate is slightly biased, with larger bias in smaller sample sizes and higher censoring rates. Further, they find that early censoring produces a more biased estimate than random or late censoring, especially for high censoring proportions. While they do not consider full cohort sizes in excess of 100 observations, they also do not consider censoring in excess of 50%, whereas in this Chapter, censoring is always in excess of 85%, reaching 99% in some circumstances.

3.6.2 Estimation of $H_0(t)$

The use of reference failure times in the simulation study design for estimation of $H_0(t)$ was imperfect, causing a degree of bias towards the end of reference times as estimates of $H_0(t)$ were carried forward. However, for comparison of relative performance of case-cohort estimators and the full cohort, this artefact is less problematic,

The simulation study shows little evidence that any considered estimator is inappropriate for estimation of $H_0(t)$ in the case-cohort design. The simulation study indicated that IPW Classic weighting offers comparable performance to Prentice Classic in estimation of $H_0(t)$, with a small reduction in bias. Post-stratification based on time provides an improvement in precision for both IPW and Prentice, with a small degree of bias-variance trade-off. This improvement in precision is greatest where sampling fractions of non-cases in the risk sets for each failure time show high variability; under Staggered Entry, with small subcohort sizes, small sampling fractions, and low non-case to case ratio. Where sampling fractions of the risk sets for each failure time show high variability, post-stratification based on time may be useful for estimation of $H_0(t)$.

3.6.3 Further Considerations

The effect of case-only risk sets on estimation of β and $H_0(t)$ was not intended to be demonstrated by the simulation study, however, it indicates that while estimates of of β are not unduly effected, estimates of $H_0(t)$ for analysis times at and following such a risk set can display profound bias, particularly with IPW weighting. Case-only risk sets are most likely to arise with staggered entry, smaller full cohorts, and smaller sampling fractions. Estimates of $H_0(t)$ and other quantities which rely on such estimates should be regarded with caution where case-only risk sets are present in the dataset.

This simulation study fails to consider the effect of early censoring on weighting methods for estimation in the case-cohort design. However, one might expect that the general patterns would follow through, with early censoring introducing increasing variability into the sampling fractions of non-cases in the risk sets for each failure time, and reducing the number of non-cases in the risk sets at later failure times in fixed entry.

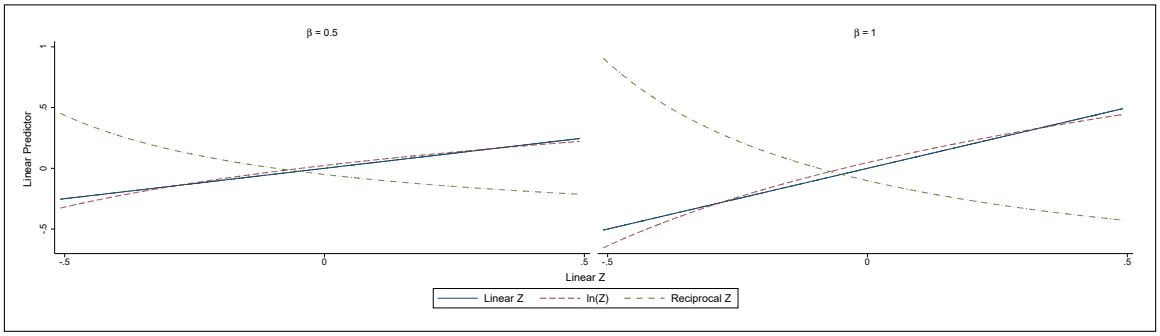
Chapter 4

Detection of Inappropriate Functional Form

4.1 Introduction

One of the key assumptions of the Cox Proportional Hazards Model is that the functional form of each covariate included in the model has a linear relationship with the hazard. That is, for a vector of covariates Z with associated coefficient vector β, the hazard function is of the form $h(t) = h_0(t)exp(\beta^T Z)$. Figure 4.1 below shows linear predictors for a single covariate $Z \sim U(0.5, 1.5)$, simulated from three functional forms; Z, ln(Z) and -1/Z (each mean-centred at 0) with $\beta = 0.5$ and $\beta = 1$, plotted against linear Z. From this figure it can be seen that the effect of inappropriate functional form depends upon the magnitude of β and the degree of difference between the "true" functional form and the candidate functional form.

Figure 4.1: Linear Predictors for Varying Functional Forms



Where a particular covariate Z_i violates the assumption of linear relationship with the hazard, and a transformation of Z_i that fulfills this assumption can be found, the solution is straightforward; the model must be fitted with Z_i transformed appropriately. However, straightforward is not always easy; to apply the above method, we require a method to detect inappropriate functional form in a fitted model.

Therneau et al. (1990) describe the use of graphical methods in the full cohort in evaluation of martingale, deviance and score residuals, where martingale residuals are plotted against covariate values to assess functional form of the covariate, deviance and martingale residuals are used to detect individual observations which are poorly predicted by the model, and score residuals are used to detect influential observations and non-proportional hazards. The purpose of this chapter is to investigate how such methods for assessing functional form of the covariate may be applied to case-cohort datasets, including the development of numerical methods by which graphical methods from individual datasets may be assessed in a simulation study over replicates.

In Section 4.2 I describe these graphical methods in the full cohort, and in Section 4.3 I describe how weighting can be applied in the case-cohort sample. In Section 4.4 I describe an analytical method using linear piecewise regression to replace subjective graphical interpretation of smooths for comparison of full cohort and case-cohort methods over replicates, and present example graphs of smooths and fits from linear piecewise regression in the full cohort and case-cohort. In Section 4.5 I perform a simulation study.

I find some evidence to suggest that weighted smooths of IPW martingale residuals can be used to assess the appropriateness of the functional form of a covariate in the case-cohort design. Power in the case-cohort design increases with sampling fraction, and with number of cases in the case-cohort sample. However, the results strongly suggest that there will be a loss of power as compared to the full cohort, and in individual datasets deviations from linearity and 0 slope may be less obvious in the case-cohort design than in a full cohort. Use of piecewise linear regression as a proxy for visual comparison of smooths in the full cohort and weighted smooths in the case-cohort appears to be inadequate, with the case-cohort substantially less sensitive to departures from linearity than the full cohort.

4.2 Full Cohort Methods

The Cox Snell residual for an observation i, with vector of covariates Z_i and associated vector of coefficients β, with entry time $t_{(0)i}$, failing or censored at time t_i is an estimate of the cumulative hazard for that observation at that time.

$$rcs_i = exp(\beta^T Z_i) \left\{ \hat{H}_0(t_i) - \hat{H}_0(t_{(0)i}) \right\}$$

The martingale residual for an observation i with failure $d_i = 0$ if the observation was censored and $d_i = 1$ if the observation failed can be defined as $rmg_i = d_i - rcs_i$. The martingale residual can be interpreted as the observed number of deaths minus the expected number of deaths, given the model, for that subject over the interval $[t_{(0)i}, t_i]$.

Where the correct functional form has been included in the model, a smooth of the martingale residuals against that covariate functional form should show zero-slope linearity - that is, the smooth should be linear with no obvious trend (Therneau et al., 1990). Note that even where the correct functional form is modelled, there may be departures from zero-slope linearity in the tails if data is sparse and there are observations with unusually large or small covariate values. (Fleming and Harrington, 2011, p184).

A number of options exist for the smooth used to assess the Martingale Residuals, including Locally Weighted Scatterplot Smoothing (LOWESS) and Kernel-Weighted Local Polynomial Smoothing (lpoly).

4.3 Case-Cohort Implementation

In the case-cohort, calculation of the individual Cox-Snell and martingale residuals does not differ from the full cohort, however, case-cohort estimates of β, $H_0(t_i)$ and $H_0(t_{(0)i})$ are used. Where case-cohort estimates are similar to the corresponding full cohort estimates, then case-cohort estimates of rmg_i should likewise be similar to full-cohort estimates.

However, interpretation of the martingale residuals in the case-cohort design is less straightforward. In a case-cohort dataset, only the subcohort sample of the non-cases is present. Since non-cases are under-represented in the case-cohort sample, the smooth of the martingale residuals must account for this under-representation in order to be interpretable. One approach to this problem is to consider each non-

case in the case-cohort dataset to be representing a number of full cohort non-cases equal to the inverse of the subcohort non-case sampling fraction, and hence weight the smooth using IPW Classic weights.

In STATA, frequency-weighted smoothing can be implemented with the `fweights` option in the local polynomial smooth command lpoly. Note that while this option is sufficient for visualising the smooth, `fweights` will provide inappropriate variances and so further options such as plotting of confidence intervals for the smooth should not be used without manual calculation of variances. For datasets containing non-integer weights, as is likely, the weight for each subject can be multiplied by some common factor, chosen to give a desired amount of precision, then rounded to the nearest integer.

STATA's LOWESS command does not allow specification of weights. However, for integer weights, the dataset can be expanded such that the number of records per subject is equal to the value of the weight for that subject, and LOWESS then performed with each record considered an individual observation. While this is possible, it is likely impractical with large datasets, since in LOWESS calculations, the number of regressions performed is equal to the number of observations.

4.4 Quantitative Assessment of Methods

Sections 4.2 and 4.3 describe methods for use in an individual dataset, where graphs are subjectively assessed by visual inspection. Subjective graphical interpretation is impractical as a method of assessment or analysis over a simulation study with a large number of replicate. Ideally, there would be an objective measure which can be used in each replicate in a simulation study to detect deviations of the martingale residual-covariate plot from zero-slope linearity and hence to allow for comparison of full cohort and case-cohort results.

One option is to record the smoothed values for each replicate and calculate the mean value of the smooth over the replications at benchmark covariate values. However, this is problematic, as while the mean smooth may show clear deviations from zero-slope linearity, this may not be apparent in the graphical visualisations from individual replicates, as would be encountered by an end-user. Similarly, the mean smooth may obfuscate deviations from zero-slope linearity that would be apparent in individual visualisations.

Where a martingale residual-covariate plot shows a linear or monotonic relation-

ship, Pearsons correlation coefficient or Spearman's rho could be used, respectively, to quantify this relationship. However, these measures are not appropriate for capture of non-monotonic relationships.

Instead, I devised replacement of the graphical assessment of good fit with a statistical test in order to assess the properties of case-cohort residuals at the level of the single dataset and fitted model. A piecewise linear regression is fitted to the martingale residuals against the candidate functional form of the covariate, with the candidate functional form reparameterised such that the regression has one linear element and a change in slope at the median value of the covariate.

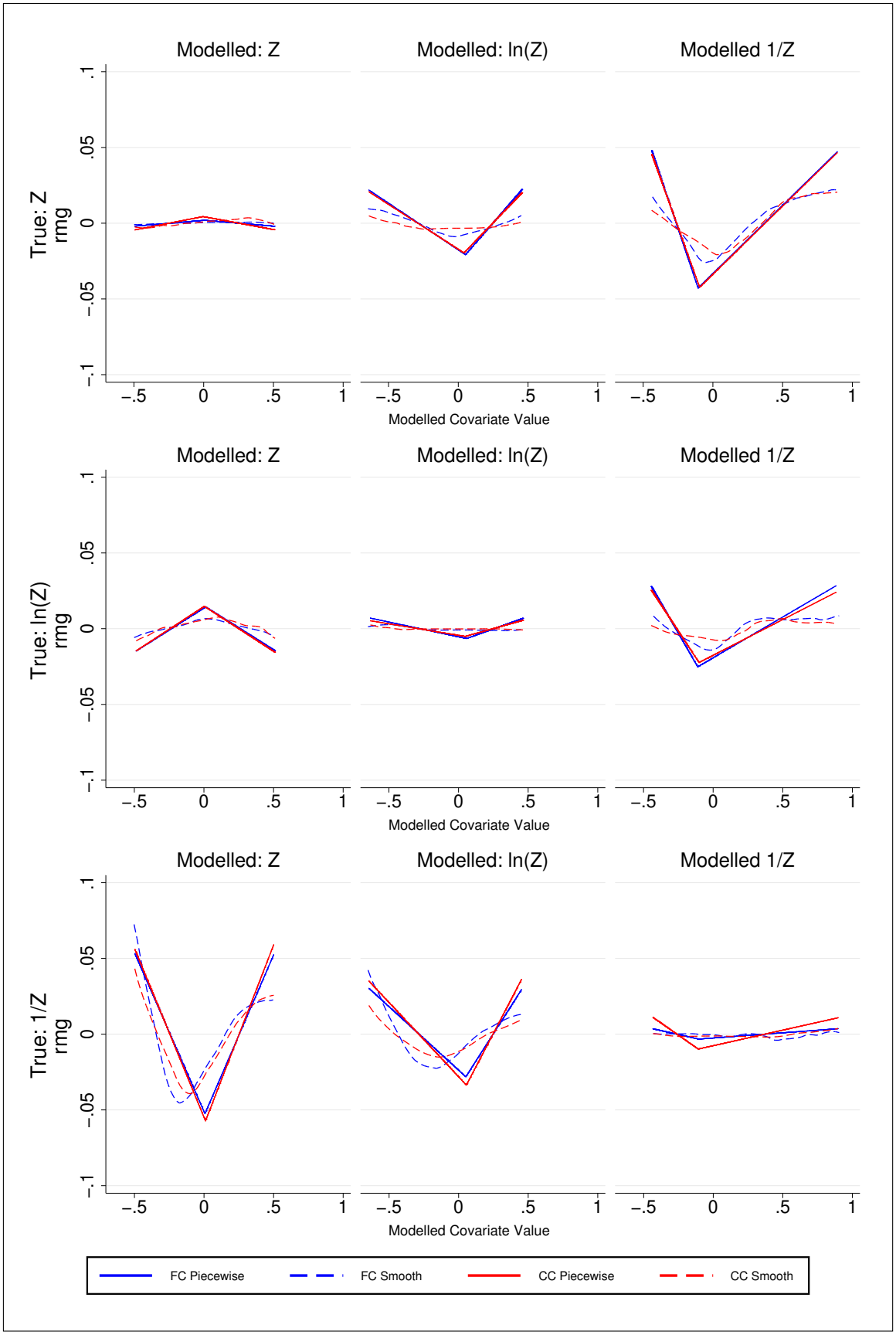
Functional forms are classified as inappropriate based on a Wald test for the marginal effects of the spline being equal to 0, i.e. whether the change in slope at the median is statistically significant. Reports from the literature (Graubard & Korn 1991) indicate that use of sampling weights in linear regressions can lead to low power for Wald tests, due to inflation of standard errors. Further, a subjective visual interpretation of a smooth may reveal departures from 0-slope linearity that do not meet statistical significance in the substitution of a piecewise linear regression for a smooth, and vice-versa, particularly when combined with the potential for low power of Wald tests in IPW-weighted linear regression. Hence it is important to note that this is not proposed as a replacement for visual inspection of smooths in individual datasets, but rather as a proxy to allow for comparison of full-cohort and case-cohort smooths in a simulation study that does not rely on individual subjective assessment of smooths.

Figures 4.2a and 4.2b show examples of full cohort and IPW-weighted local polynomial smooths and linear piecewise regressions of martingale residuals against candidate functional forms for staggered entry, subcohort size 1000 and $\beta = \ln(3)/\text{SD}$, generated according to the data generating mechanism in Section 4.5. Smooths use the STATA defaults Epanechnikov kernel, degree 0, smooth at min(N, 50) points, and rule-of-thumb bandwidth. Sampling fraction 15% with non-case to case ratio 1:1, and sampling fraction 3% with non-case to case ratio 4:1 are shown

In the 15% sample, the case-cohort smooths and piecewise regressions appear similar to those of the full cohort. Additionally, the correctly modeled form appears to have smooths and linear fits closer to 0-slope linearity than the incorrectly modeled forms. Coefficients for the regression are substantially smaller in the 3% sample, and behaviour of smooths and piecewise regressions are less similar between the full cohort and the case cohort.

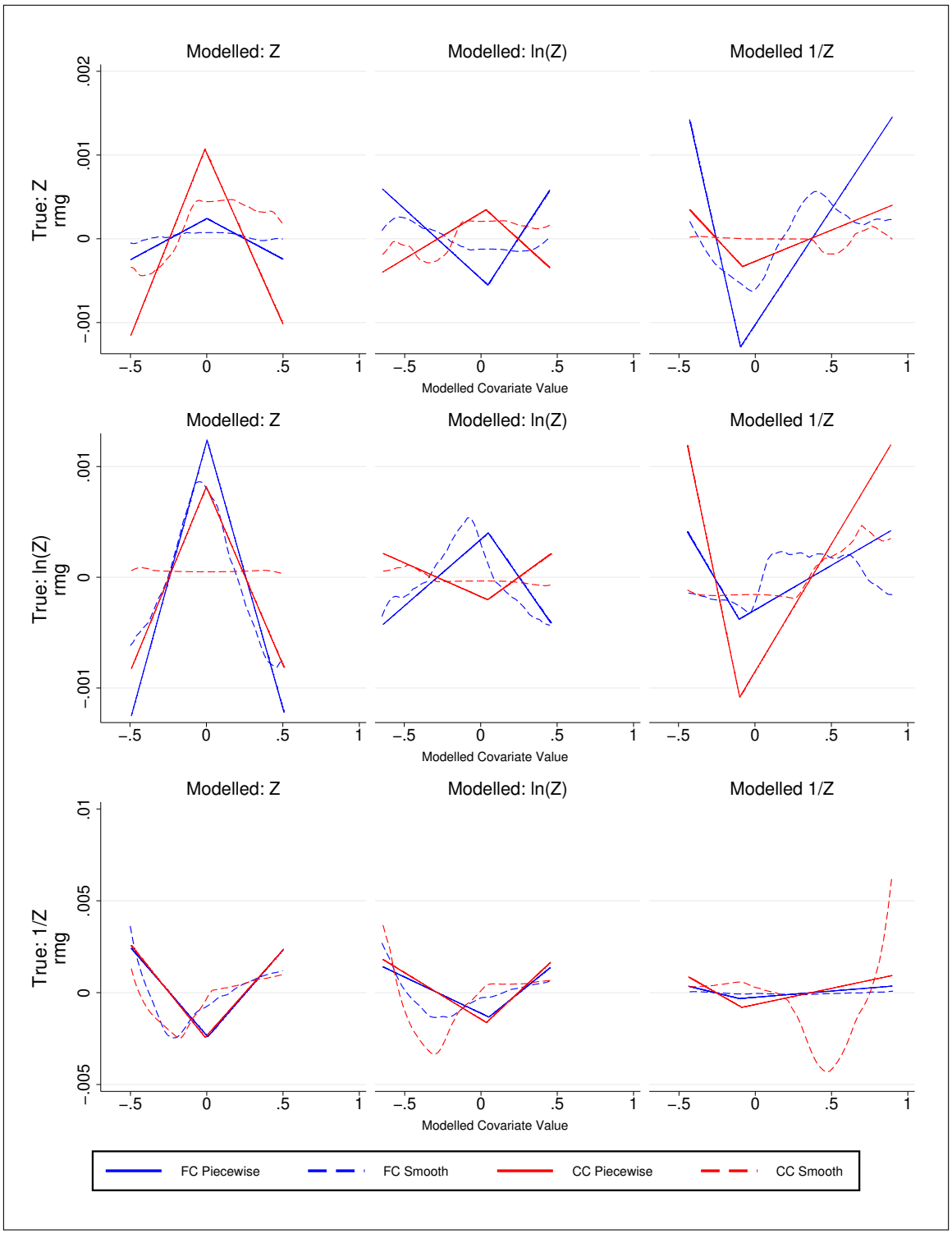
Figure 4.2: Example Piecewise Linear Fits and Local Polynomial Smooths

(a) sampling fraction 15%; non-case to case ratio 1:1 - y-axis: martingale residuals value



FC=Full Cohort; CC = Case-Cohort

(b) sampling fraction 3%; non-case to case ratio 4:1 - y-axis: martingale residuals value



FC=Full Cohort; CC = Case-Cohort

4.5 Simulation Study

The purpose of this simulation study is to compare the results of the piecewise regression test method in the full cohort the case-cohort as a measure of the adequacy of weighted graphical assessment of IPW-weighted martingale residuals for functional form in the case-cohort.

4.5.1 Data Generating Mechanism

Data generation was carried out as described in Chapter 2, with the following specifications:

The presence of particularly large or small covariate values can have a profound effect on the appearance of a smooth and the significance of marginal splines, particularly under certain transformations. While in individual plots a subjective judgement can be made on which values can be considered outliers, for the purposes of a simulation study, such effects should be minimized insofar as possible. As such, full cohorts were simulated with six initial continuous covariates independently drawn from $\sim U(0.5, 1.5)$. For simulation of survival times, two covariates remained untransformed, two covariates were transformed to log form, and two covariates were transformed to reciprocal form, resulting in three pairs of covariates. Each covariate was centred at mean 0 prior to generation of survival times.

Survival times were generated with vector of coefficients for each covariate pair equal to hazard ratios per standard deviation of 2 and 3, respectively. 1 million initial observations of each functional form were generated to estimate standard deviations which were 0.29, 0.31, and 0.36 for linear, log, and reciprocal forms, respectively. Distributions of mean-centred covariates and their linear predictors are demonstrated in Figure 4.3.

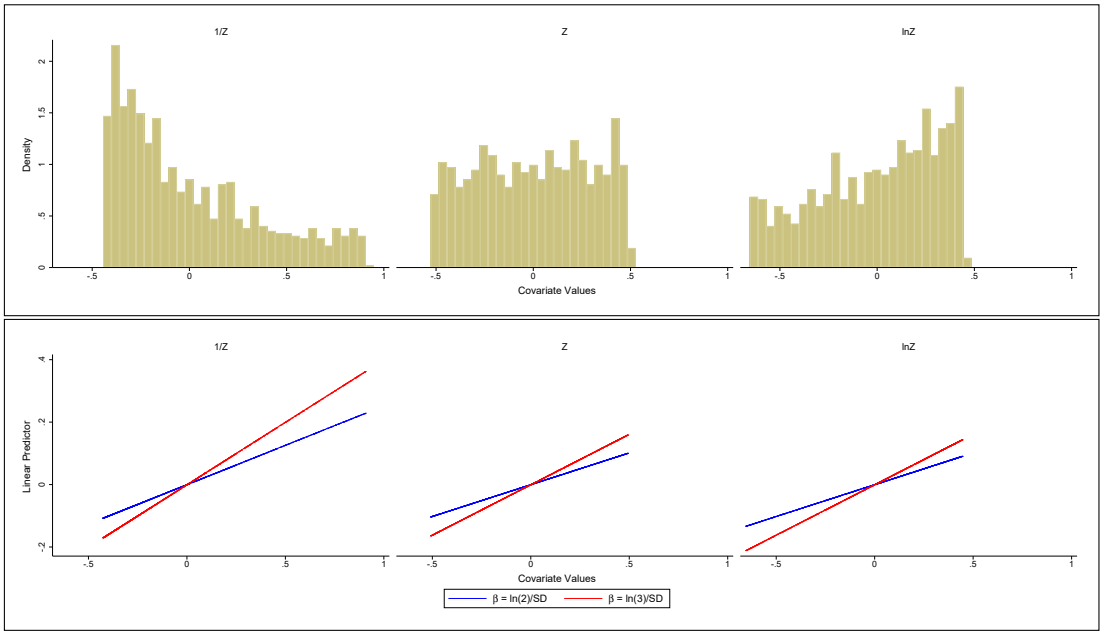
4.5.2 Estimands

For each of the candidate functional forms linear, log and reciprocal, IPW Classic weighted Cox models were fitted with all six covariates included in the model, transformed to the candidate functional form, and martingale residuals were calculated. Martingale residuals for the full-cohort were also calculated.

IPW Classic was used to weight the data when creating the splines and performing the regression in the case-cohort sample. Wald and robust Wald test results for the marginal effect of the splines were recorded in the full cohort and case-cohort,

respectively. A number of methods for estimation of standard errors, including Huber sandwich, jacknife and bootstrap were considered, but exploratory analysis indicated that all methods gave similar power in the case-cohort linear regression. Robust (Huber sandwich) standard errors were used in the simulation study due to their smaller computational time.

Figure 4.3: Distributions and Linear predictors of Covariate Functional Forms



4.5.3 Methods

Full cohort and IPW Classic-weighted case-cohort methods were considered. Martingale residuals are calculated based on estimates of $H_0(t)$ and β. Given the results of Chapter 3, it was not expected that Prentice weighting would result in large differences in martingale residuals compared to IPW, except possibly in the presence of case-only risk sets, which can lead to inappropriate estimates of $H_0(t)$.

Exploratory simulations were carried out such that where case-only risk-sets were present in the datasets, martingale residuals were calculated with estimates for $H_0(t)$ for such observations replaced by the previous non-case only risk set estimate of $H_0(t)$. However, this modification did not appear to affect the results and hence was not included in the full simulation study.

4.5.4 Performance Measures

A cutoff criteria of $p < 0.05$ was used to classify the Wald tests as indicating inappropriate functional form. MCSE bounds for power and Type 1 error were calculated using the `simsum` STATA package.

4.5.5 Results

Results for staggered entry and fixed entry were similar. Hence, for clarity and brevity only staggered entry is presented here in Tables **??**, 4.2, and 4.3.

In the full cohort, power is primarily influenced by the number of cases, β and the relationship between the true and candidate functional form. Full cohort size has a minimal effect on power in this simulation study. Power is greatest for distinguishing between linear and reciprocal forms, and lowest for distinguishing between log and linear forms.

In the case cohort, β and the relationship between the true and candidate functional form have a similar influence on power as in the full cohort. Power increases with sampling fraction. Power increases with non-case to case ratio and subcohort size, but where this results in the number of cases being similar, power is also similar within a particular sampling fraction (e.g. subcohort size 200, noncase to case ratio 1:1 vs subcohort size 1000, noncase to case ratio 4:1). Loss of power is substantial, particularly at the 3% sampling fraction, however. This loss of power tends to be associated with a lower Type 1 error rate than the full cohort.

Overall, results show that, as evaluated by piecewise linear regression, the case-cohort is substantially less sensitive to departures from linearity than the full cohort, with this reflected in both power and Type 1 error rates.

Table 4.1: Classification of Inappropriate Functional Form Rates for True Form Z, Staggered Entry

HR/SD	N^{SC}	$\frac{N^{NC}}{N^C}$	α	N^C	N	Modelled:	Z				ln(Z)				1/Z			
						Sample:	FC		IC		FC		IC		FC		IC	
							Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE
2	200	4	3	50	6667		3.1	0.5	1.9	0.4	5.7	0.7	2.0	0.4	13.6	1.1	4.3	0.6
			15	48	1333		2.4	0.5	2.1	0.5	5.2	0.7	3.4	0.6	13.8	1.1	7.1	0.8
		1	3	194	6667		3.4	0.6	4.0	0.6	15.5	1.1	4.7	0.7	51.8	1.6	10.9	1.0
			15	174	1333		3.3	0.6	3.7	0.6	16.6	1.2	7.5	0.8	51.5	1.6	21.1	1.3
	1000	4	3	248	33333		2.8	0.5	1.3	0.4	15.5	1.1	2.5	0.5	58.9	1.6	13.4	1.1
			15	241	6667		2.3	0.5	2.5	0.5	18.1	1.2	8.5	0.9	62.9	1.5	28.7	1.4
		1	3	971	33333		3.2	0.6	1.8	0.4	61.8	1.5	9.3	0.9	99.9	0.1	34.8	1.5
			15	870	6667		3.8	0.6	4.1	0.6	64.4	1.5	26.2	1.4	99.8	0.1	80.6	1.3
3	200	4	3	50	6667		0.5	0.2	0.4	0.2	4.2	0.6	0.8	0.3	12.7	1.1	2.4	0.5
			15	48	1333		1.4	0.4	0.9	0.3	3.9	0.6	1.5	0.4	15.0	1.1	5.0	0.7
		1	3	194	6667		1.4	0.4	0.1	0.1	14.1	1.1	1.6	0.4	65.2	1.5	6.6	0.8
			15	174	1333		1.4	0.4	1.9	0.4	21.3	1.3	6.9	0.8	73.9	1.4	29.5	1.4
	1000	4	3	248	33333		0.5	0.2	0.0	0.0	15.3	1.1	1.5	0.4	65.6	1.5	9.3	0.9
			15	241	6667		0.5	0.2	0.5	0.2	20.0	1.3	4.0	0.6	77.9	1.3	32.0	1.5
		1	3	971	33333		1.5	0.4	0.2	0.1	76.0	1.4	4.7	0.7	100.0	0.0	37.4	1.5
			15	870	6667		1.6	0.4	1.9	0.4	85.9	1.1	33.3	1.5	100.0	0.0	95.4	0.7

FC= Full cohort; IC = IPW Classic; N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio; α sampling fraction (%); N Full cohort size;

Table 4.2: Classification of Inappropriate Functional Form Rates for True Form ln(Z), Staggered Entry

HR/SD	N^{SC}	$\frac{N^{NC}}{N^C}$	α	N^C	N	Modelled:	Z				ln(Z)				1/Z			
						Sample:	FC		IC		FC		IC		FC		IC	
							Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE
2	200	4	3	50	6667		4.4	0.6	2.4	0.5	2.5	0.5	1.3	0.4	5.0	0.7	1.5	0.4
			15	48	1333		6.5	0.8	3.8	0.6	3.8	0.6	2.6	0.5	6.6	0.8	3.3	0.6
		1	3	194	6667		14.1	1.1	4.5	0.7	3.7	0.6	2.1	0.5	14.2	1.1	3.5	0.6
			15	174	1333		13.0	1.1	7.9	0.9	3.7	0.6	3.5	0.6	15.3	1.1	7.3	0.8
	1000	4	3	248	33333		14.1	1.1	2.5	0.5	3.4	0.6	1.0	0.3	15.7	1.2	4.4	0.6
			15	241	6667		19.5	1.3	7.0	0.8	3.3	0.6	1.6	0.4	17.7	1.2	7.4	0.8
		1	3	971	33333		55.1	1.6	6.9	0.8	2.5	0.5	1.5	0.4	63.5	1.5	9.9	0.9
			15	870	6667		56.2	1.6	19.4	1.3	3.3	0.6	3.6	0.6	61.9	1.5	24.3	1.4
3	200	4	3	50	6667		2.3	0.5	1.1	0.3	1.0	0.3	0.7	0.3	2.2	0.5	0.8	0.3
			15	48	1333		3.4	0.6	1.8	0.4	1.2	0.3	0.6	0.2	4.6	0.7	1.5	0.4
		1	3	194	6667		12.1	1.0	1.7	0.4	1.6	0.4	0.6	0.2	15.3	1.1	1.5	0.4
			15	174	1333		20.4	1.3	7.0	0.8	2.1	0.5	1.4	0.4	17.4	1.2	5.2	0.7
	1000	4	3	248	33333		14.6	1.1	1.6	0.4	1.6	0.4	0.1	0.1	14.1	1.1	1.7	0.4
			15	241	6667		19.0	1.2	4.4	0.6	1.0	0.3	1.1	0.3	18.5	1.2	5.3	0.7
		1	3	971	33333		73.6	1.4	4.6	0.7	1.1	0.3	0.7	0.3	79.1	1.3	5.9	0.7
			15	870	6667		80.1	1.3	28.8	1.4	2.6	0.5	1.9	0.4	84.1	1.2	30.9	1.5

FC= Full cohort; IC = IPW Classic; N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio; α sampling fraction (%); N Full cohort size;

Table 4.3: Classification of Inappropriate Functional Form Rates for True Form 1/Z, Staggered Entry

HR/SD	N^{SC}	$\frac{N^{NC}}{N^C}$	α	N^C	N	Modelled:	Z				ln(Z)				1/Z			
						Sample:	FC		IC		FC		IC		FC		IC	
							Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE	Class	MCSE
2	200	4	3	50	6667		13.7	1.1	3.9	0.6	4.6	0.7	2.4	0.5	1.5	0.4	1.1	0.3
			15	48	1333		15.1	1.1	8.0	0.9	5.7	0.7	4.8	0.7	2.1	0.5	3.2	0.6
		1	3	194	6667		58.1	1.6	11.7	1.0	17.0	1.2	4.3	0.6	1.6	0.4	1.8	0.4
			15	174	1333		53.4	1.6	25.6	1.4	17.9	1.2	9.2	0.9	2.5	0.5	3.3	0.6
	1000	4	3	248	33333		68.1	1.5	15.4	1.1	20.7	1.3	4.8	0.7	2.0	0.4	0.7	0.3
			15	241	6667		68.9	1.5	33.3	1.5	21.3	1.3	8.7	0.9	1.7	0.4	2.4	0.5
		1	3	971	33333		99.9	0.1	42.7	1.6	72.4	1.4	11.3	1.0	2.9	0.5	1.5	0.4
			15	870	6667		99.8	0.1	83.5	1.2	72.4	1.4	33.3	1.5	2.8	0.5	4.0	0.6
3	200	4	3	50	6667		13.3	1.1	1.8	0.4	2.7	0.5	0.3	0.2	0.3	0.2	0.1	0.1
			15	48	1333		15.7	1.2	5.1	0.7	4.1	0.6	1.3	0.4	0.9	0.3	1.1	0.3
		1	3	194	6667		70.8	1.4	7.1	0.8	16.2	1.2	1.6	0.4	0.4	0.2	0.2	0.1
			15	174	1333		80.7	1.2	39.4	1.5	26.2	1.4	8.8	0.9	0.9	0.3	1.4	0.4
	1000	4	3	248	33333		75.2	1.4	8.4	0.9	15.8	1.2	0.8	0.3	0.4	0.2	0.0	0.0
			15	241	6667		84.4	1.1	38.8	1.5	22.5	1.3	6.1	0.8	0.4	0.2	0.2	0.1
		1	3	971	33333		100.0	0.0	46.9	1.6	85.8	1.1	3.9	0.6	0.6	0.2	0.0	0.0
			15	870	6667		100.0	0.0	98.9	0.3	92.4	0.8	52.3	1.6	1.5	0.4	1.3	0.4

FC= Full cohort; IC = IPW Classic; N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio; α sampling fraction (%); N Full cohort size;

4.6 Discussion

Evaluation of the behaviour of smooths of martingale residuals in the case-cohort design presents a number of challenges. Visual assessment of individual smooths in each simulated dataset of the simulation study is both impractical and subjective. However, the statistical test based on piecewise regression is also imperfect. While the goal of the piecewise linear regression was not to serve as a formal statistical test for use in individual datasets, it was hoped that it would serve as a useful proxy for comparison of smooths in the full cohort and weighted smooths in the case-cohort. However, the method appears to be inadequate.

Low power of Wald tests in weighted linear regression where there are unequal sampling fractions is a known issue in complex survey sampling, with no clear solution obvious in the literature. Resolution of this issue is, further, beyond the scope of this thesis. It is hence difficult to distinguish between loss of power due to the case-cohort design itself, and loss of power due to the inadequacy of Wald tests in the linear regression. The correspondence in the case-cohort of low power to low Type 1 error rate indicates that a degree of the loss of power may be attributed to the known issue of low power of Wald tests in IPW-weighted linear regression due to inflation of standard errors. However, the degree of power loss suggests that other factors may be at play.

Given the results of Chapter 3, it appears unlikely that the loss of power is made up in any large part by differing estimates of β and $H_0(t)$ in the full cohort and the case cohort. The loss of power here is far in excess of that which might be expected from the results for estimation of β seen in Chapter 3. Note, however, that this comparison is not like-with-like. Chapter 3 assesses power for estimation of a Cox model, whereas this chapter assesses power for a linear regression.

The coefficients for the linear piecewise regressions are substantially smaller in the worst-performing scenarios than the best-performing scenarios, as seen in the differing scales in Figure 4.2a. However, the magnitude of the coefficients does not differ substantially between the full cohort and the case cohort, despite the large differences in power.

It is possible that the piecewise linear regression is a worse approximation to the smooth in the case-cohort than in the full-cohort. It is also possible that the approximations perform similarly, but that the weighted smooths themselves differ significantly between the full cohort and the case cohort. However, from exploratory

visual inspection of a sample of smooths, this does not appear to account for the degree of loss of power.

Hence, the results of this simulation study are more advisory and indicative than they are prescriptive.

There is no evidence from the results that IPW-weighted smooths of the martingale residuals gives inappropriate Type 1 error. The simulation study gives some evidence to suggest that, when properly weighted, smooths of the IPW martingale residuals can be used to assess the appropriateness of the functional form of a covariate in the case-cohort design. However, the results strongly suggest that there will be a loss of power as compared to the full cohort. As such, in individual datasets deviations from linearity and 0 slope may be less obvious in the case-cohort design than in a full cohort. Power in the case-cohort design increases with sampling fraction, and with number of cases in the case-cohort sample, with subcohort size and non-case to case ratio of lesser influence.

Interestingly, in exploratory studies, modifications for case-only risk sets did not affect the results. This is possibly due to the fact that estimates of $H_0(t)$ made only a small contribution to the martingale residuals in the simulation study, due to the relative magnitudes of β and $H_0(t)$. In individual datasets, when assessing a graphical smooth, guidance already indicates that the presence of particularly large or small covariate values can have a profound effect on the appearance of a smooth, and such values should be treated with caution. Where case-only risk sets are present in the data, caution would indicate that the smooth should also be assessed for martingale residuals modified to exclude case-only risk sets from the estimates of $H_0(t)$.

Chapter 5

Detection of Non-Proportional Hazards

5.1 Introduction

A critical assumption of the Cox proportional hazards model, so critical that it is included in the name, is that the hazards are proportional over time. i.e. that covariates are multiplicatively related to the hazard. In the full cohort, a number of methods exist for detection of violations of the assumption of proportional hazards. In Section 5.2 I describe a selection of these methods in the full cohort. In section 5.3, I describe their implementation in the case-cohort design. In Section 5.4 I perform a simulation study to assess the performance of Schoenfeld residuals, scaled Schoenfeld residuals, and inclusion of time-varying covariates in the model for detection of non-proportional hazards in the case-cohort design. I find that where risk set sizes are not overly variable, all three methods are appropriate for use in the case-cohort design, with similar power. Where case-cohort risk set sizes are more variable, methods based on Schoenfeld residuals and scaled Schoenfeld residuals show high Type 1 error rate.

5.2 Full Cohort Methods

In the Cox model, there are three general classes of approach for assessment of non-proportional hazards; graphical interpretation of survival curve estimates, statistical tests of residuals, and model-based statistical tests. In the following sections I outline three survival curve methods, log-log plots, Kaplan Meier baseline survival estimate plots, and comparison of Kaplan Meier and Cox-predicted baseline survival curves; two statistical tests based on residuals, correlation of Schoenfeld residuals with time, and Grambsch and Therneau's scaled Schoenfeld residual test statistics; and two

model-based statistical tests, Cox's interaction method, and Schemper's piecewise regression method. These methods were chosen based on prevalence in the literature and ease of implementation in STATA. In addition to these methods, a number of further methods are extant in the literature, e.g. the score test of Lin et al. (2006), tests based on cumulative sums of martingale residuals Lin et al. (1993), Therneau's score test (Therneau et al., 1990) and Moreau's score test (Moreau et al., 1985).

5.2.1 Graphical Interpretation of Survival Curve Estimates

Graphical methods for detection of non-proportional hazards in the full cohort based on survival curves include log-log plots, Kaplan Meier baseline survival estimate plots, and comparison of Kaplan Meier and Cox-predicted baseline survival curves.

Log-log plots, first suggested by Kalbfleisch-Prentice (1980), plot $-ln(-ln(\hat{S}(t)))$ against $ln(t)$ for each category of a nominal or ordinal covariate, where $\hat{S}(t)$ is the estimated survival function based on the Cox model. When the curves are not parallel it indicates violation of the proportional hazards assumption.

Kaplan and Meier (1958) define a nonparametric maximum likelihood estimate of the survivor function as $\hat{S}(t) = \prod_{t_{(j)} \leq t}(\frac{N_j - d_j}{N_j})$ for N_j = the number at risk of failure just before $t_{(j)}$ and d_j = the number of failures at $t_{(j)}$. Where Kaplan Meier baseline survival estimate curves for each category of a nominal or ordinal covariate against time cross or converge, it indicates violation of the proportional hazards assumption.

Divergence of observed Kaplan Meier and Cox-predicted baseline survival curves indicates violation of the proportional hazards assumption.

5.2.2 Formal Statistical Tests

5.2.2.1 Tests based on Residuals

Schoenfeld (1982) described partial residuals for the proportional hazards model. Schoenfeld residuals are defined for each event or failure, with a seperate Schoenfeld residual for each covariate in the model. For a particular observation and particular covariate, the Schoenfeld residual is the difference between the value of the covariate and its weighted mean conditioned upon the risk set at the failure time of that observation. Using the risk set notation earlier described, in the full cohort, Schoenfeld residuals for covariate Z_k and an observation failing at time $t_{(j)}$ can be defined as:

$$rs = Z_{[j]k} - E\left(Z_{[j]k}|R_{(j)}\right)$$

where:

$$E\left(Z_{[j]k}|R_{(j)}\right) = \frac{exp(\hat{\beta}Z_{[j]})Z_{[j]k} + \sum_{l \in R_{(j)}} exp(\hat{\beta}Z_l)Z_k}{exp(\hat{\beta}Z_{[j]}) + \sum_{l \in R_{(j)}} exp(\hat{\beta}Z_l)}$$

Under the proportional hazards assumption, the Schoenfeld residuals have mean zero and are uncorrelated with time. Graphical assessment and formal correlation tests of this relationship can be employed to assess violations of the assumption.

Grambsch and Therneau (1994) present scaled Schoenfeld residuals $rsc_{[j]}$. Let V refer to the variance of the vector of coefficients β. Grambsch and Therneau (1994) originally proposed scaling the Schoenfeld residuals by the weighted variance of the covariates at each failure time time $t_{(j)}$. They note that this weighted variance can become unstable as the number of observations in the risk set decreases, they propose substitution with the average variance V/N^C, such that scaled Schoenfeld residuals are estimated as

$$rsc_{[j]} = N^C \hat{V}(\beta)^{-1} rs'_{[j]}$$

Under the proportional hazards assumption, smoothed scaled Schoenfeld residuals can be interpreted as a nonparametric estimate of the log hazard-ratio function and should have slope 0 when plotted against functions of event time. Grambsch & Therneau present a test statistic based on the least squares slope of linear regressions of scaled Schoenfeld residuals against time for individual covariates, with this test statistic asymptotically distributed as χ^2 with degrees of freedom 1. They also present a global test for m covariates, with the test statistic asymptotically distributed as χ^2 with degrees of freedom m.

5.2.2.2 Model-Based Tests

Non-proportional hazards implies that covariates will have different impacts on the hazard rate at different analysis times. Schemper (1992) suggests splitting analysis time at some predetermined value(s), fitting separate Cox regressions in each element of the partition, and examining whether parameter estimates differ between Cox regressions in the different subsets of time.

Cox (1972) suggests adding a time-varying covariate to the model in the form of an interaction between a function of time and the covariate of interest. The significance of this interaction can then be assessed by Wald, Likelihood ratio or score tests.

5.2.3 Comparison of Methods

Log-log plots have come under substantial criticism (Chastang, 1983; Schemper, 1992) for failure to consistently and correctly detect nonproportionality. Assessment of parallel curves and divergence of curves for log-log plots and comparison of Kaplan-Meier and Cox-predicted baseline survival curves is a subjective graphical assessment, with an inherent lack of objectivity. Assessment of log-log plots and crossing of Kaplan Meier estimates is straightforward only for categorical or ordinal covariates, with continuous covariates requiring some decision on partitioning of the covariate. Convergence of Kaplan Meier estimates is also subjectively assessed.

Of the formal statistical tests described above, Cox's method is most computationally intensive. All require a choice of the function of time $g(t)$ for which non-proportionality will be assessed. Analysis time, log of analysis time, and rank of analysis time are popular choices, and are often native options in statistical software. Park and Hendry (2015) note that choice of function of time can have a profound effect on the performance of Grambsch and Therneau's scaled Schoenfeld residual tests, and recommend choosing rank of time where outliers are present in the dataset. Schemper's method also requires a choice of how to partition the time-scale.

A number of comparisons of methods for detection of NPH in the full cohort are extant in the literature, including Austin (2018), Grant et al. (2014), Song and Lee (2000), Ng'andu (1997) and Hess (1995). Broadly, the literature suggests that relative performance of tests and power of tests depends upon the form and magnitude of departure from the proportional hazards assumption, correlation between covariates, covariate distributions, and the number of cases observed in the dataset. While the form and magnitude of the violation of proportion hazards will not be known outside of simulation studies, the other factors can be assessed from the data.

A distinction can be drawn between methods that detect the presence of non-proportional hazards in the model as a whole, and methods that allow for the identification of which covariates(s) display non-proportional hazards - that is, global and covariate-specific methods. If the goal is only to assess whether non-proportional hazards are present, global tests are sufficient. However, should the goal be to allow for analysis under the Cox model, perhaps by means of inclusion of an interaction with time in the model, or stratification, identification of the specific covariate(s) that display non-proportional hazards is necessary.

Amongst the graphical methods, log-log plots and crossing Kaplan-Meier curves are

covariate-specific, whereas comparison of Kaplan Meier and Cox-predicted baseline survival curves is a global method. Linear correlation of Schoenfeld residuals with $g(t)$ are covariate-specific, whereas Grambsch & Therneau provide both covariate-specific and global tests. Piecewise regression and the extended Cox model can be implemented as covariate-specific or global tests, depending on which covariates are allowed to interact with time, and the use of single or multiple parameter Wald tests. Where non-proportional covariates are correlated with proportional covariates, one might expect that covariate-specific tests and one-at-a-time inclusion of covariates interacting with time would lead to high Type 1 error rate. However, single-parameter Wald tests following the global inclusion of covariates interacting with time via Cox or Schemper's methods may allow for identification of the specific covariate(s) that display non-proportional hazards in the presence of correlation.

Winnett and Sasieni (2001) note that for scaled Schoenfeld residuals, the substitution of the average variance for the weighted variance of the covariates at each failure time may result in misleading estimates of time-varying coefficients when variance of covariates changes substantially over time.

5.3 Case-Cohort Adaptations

5.3.1 Graphical Interpretation of Survival Curve Estimates

The survival function $S(t)$ can be defined as the exponent of the negative cumulative hazard function $H(t)$. The survival function can therefore be estimated as

$$\hat{S}(t) = exp(-\hat{H}_0(t)exp(\hat{\beta}^T Z))$$

It is logical, therefore, to expect that the methods from Chapter 3 regarding estimation of $H_0(t)$ and β will extend to the estimation of $S_0(t)$ and $S(t)$. In the context of absolute risk $(1 - S(t))$ Sanderson et al. (2013) estimate $S(t)$ in the case-cohort design with Prentice Classic weights and find that on average the absolute risk tends to be overestimated at low subcohort sampling fractions, and the variability at low subcohort sampling fractions is also greater. These results also correspond to the findings from Chapter 3 of this thesis. Recall, however, from Chapter 3, that empirical standard error of estimates of $H_0(t)$ is greater towards the beginning of analysis time, and that presence of case-only risk sets can introduce substantial bias. Diversions from the patterns expected from datasets with proportional hazards should be regarded with caution when they appear only early in analysis time or as the result of case-only risk sets being present in the data.

5.3.2 Methods Based on Residuals

In the case-cohort design, the Schoenfeld residuals differ from the full cohort only insofar that case-cohort estimates of β are used, and as regards the risk sets on which the mean of the covariate is conditioned. Let Prentice-weighted Schoenfeld residuals refer to residuals calculated with all observations taking weight equal to 1, and non-subcohort cases considered at risk only at their failure time. Let IPW-weighted Schoenfeld residuals refer to residuals calculated with all cases taking weight equal to 1, and subcohort non-cases taking weight appropriate to the IPW variant (e.g $\frac{N^{NC}}{N^{*NC}}$ for IPW Classic).

Where small full cohort risk sets are present in the data, due to random chance, covariates in individual subcohort risk sets are at greater risk of being unrepresentative of the covariates in the full cohort risk sets, with this effect exacerbated by small sampling sizes. Tests based on correlation can be sensitive to outliers (Abdullah, 1990) and this may affect the performance of tests based on correlation of Schoenfeld residuals with time.

In their critique of the use of the average variance substitution for scaled Schoenfeld residuals, Winnet & Sasieni note that this substitution may lead to inappropriate residuals when effect size is large or covariates have skewed distributions. However, one would also expect that the weighted variance of the covariates at each failure time $t_{(j)}$ would show increasing variation where risk set sizes vary. In the case cohort, additional variation in risk set sizes is introduced by random sampling of the full cohort. Under IPW weighting, the inclusion of all cases and a weighted sample of the non-cases could also lead to increased variation in scenarios where, due to random chance, the non-cases sampled from particular risk sets are not representative of the non-cases in the full cohort risk set, with this then exacerbated by weighting. Under Prentice weighting, non-cases are only included in the risk set at their failure times, so this effect may be reduced. The combination of the average variance substitution with the possibility of non-representative case-cohort risk sets impacting on calculation of the Schoenfeld residuals could lead to an additive issue with use of scaled Schoenfeld residuals tests in the case-cohort design.

Note that since Schoenfeld and scaled Schoenfeld residuals are defined only for cases, and the test statistic for scaled Schoenfeld Residuals relies only upon the failure times, whether Prentice or IPW weighting was used to calculate the residuals is not relevant for performing the tests. For scaled Schoenfeld residuals, care must be taken with the variance used for calculation of the residuals and implementation

of the statistical tests. The average variance as defined by Gramsch & Therneau is the inverse of the information matrix, divided by the total number of cases. Hence, the robust variance, as used for estimation of coefficients, is inappropriate and the model-based variance should be used.

Xue et al. (2013) perform a simulation study that assesses performance of Prentice-Weighted Schoenfeld Residuals in the case-cohort design. They use 1000 replicates of a full cohort of 2000 with a random subcohort of 500 subjects and a uniform censoring distribution such that the event rate was set to be between 5-10%. They state that "Several different cohort and subcohort sample sizes were assumed, however ... changes in sample size did not affect the findings" however they do not describe these full cohort and subcohort sizes. Analysis time-scale is not explicitly described, but it appears to be fixed entry. They assess models with a single binary covariate, a single continuous covariate generated from a standard normal distribution, and both a binary and an independent continuous variable. They assess the PH assumption by calculating a Pearson correlation coefficient and its significance for the covariate between its Schoenfeld residuals and each of time, rank order of time, and Kaplan Meier estimates. Broadly speaking, all tests showed similar Type 1 error, within 1% of a nominal 5%. Power was also similar, with analysis time showing somewhat reduced power for the single binary covariate with certain forms of non-proportional hazards. However, note that the simulation study described had a subcohort sampling fraction of 25% and 500 subcohort members, with a maximum of 200 cases at the 10% event rate. If the assumption of fixed entry is correct, the minimum risk set size would therefore be 301 subjects for the last failure. It is therefore unclear whether the results of this simulation study would apply to data with smaller risk set sizes, at might occur with smaller full cohorts, smaller sampling fractions or data analysed with age as time-scale.

5.3.3 Model-Based Tests

The model-based tests described above are, essentially, tests of the power of the model to detect an interaction effect. While Chapter 3 did not consider interactions specifically, one might expect to see similar results as were demonstrated for estimation of β in that chapter. No obvious further necessary modification for the case-cohort design presents itself.

5.4 Simulation Study

The purpose of this simulation study is to compare full cohort statistical tests for detection of non-proportional hazards with case-cohort methods.

5.4.1 Data Generating Mechanism

Data generation was carried out as described in Chapter 2, with the following specifications:

One independent binary covariate was generated from $Binomial(1, 0.5)$. One independent Normal covariate was generated from $Normal(0, 1)$. Two binary covariates with correlation 0.5 were generated from $Binomial(1, 0.5)$.

Survival times were generated with the normal covariate and one correlated binary covariate displaying non-proportional hazards. For each combination of subcohort size, non-case to case ratio, sampling fraction, and time-scale, initial reference proportional hazards datasets were generated with full cohorts of size 10,000 and coefficients for all four covariates equal to ln(2)/SD. The changes in hazard ratio Δ from the 25th percentile to the 75th percentile of survival times of cases were 0.8 and 1.25. To assess Type 1 error for global tests, full cohorts were also generated with $\Delta = 1$, i.e. with no violations of the proportional hazards assumption. Examples of the associated values of β_{phi} and ϕ are detailed in Chapter 2, Table 2.4. Example graphs of Hazard ratios against time are shown in Chapter 2, Figure 2.2.

5.4.2 Target

For all methods investigated in this study, the target was evaluation of the null hypothesis of proportional hazards, as assessed by the various methods.

5.4.3 Methods

Full Cohort, IPW-Classic and Prentice-weighted methods were considered. In exploratory simulations, use of time-dependent weights as for estimation of β or Prentice Classic as for estimation of $H_0(t)$ in Chapter 3, and removal of case-only risk sets, had only minimal effects on results. Hence they were not considered.

In exploratory simulations, choice of function of time from analysis time, log of analysis time and rank of analysis time had minimal impact on the relative performance of tests in the case-cohort as compared to the full cohort. Hence, only analysis time was considered.

The methods investigated in this chapter can be classified as single-parameter or global tests. Note that while the single-parameter Cox test implemented here accounts for non-proportional hazards in the other covariates, the single-parameter tests for Grambsch & Therneau and correlation of Schoenfeld residuals with time do not.

Single-Parameter Tests

1. Pearson correlation of Schoenfeld residuals with $g(t)$.

2. Cox's method, assessed for the interaction of a single covariate with $g(t)$ following inclusion of interactions of all covariates with $g(t)$ in the model.

3. Grambsch & Therneau's method, assessed for a single parameter

Global Tests

1. Cox's method, assessed for the interactions of all covariates in the model with $g(t)$.

2. Grambsch & Therneau's method, assessed for all covariates in the model.

For Cox's methods, Wald tests were used to assess the significance of the interactions with $g(t)$.

5.4.4 Performance Measures

Performance measures for each target were power and Type 1 error, with a a cutoff of $p = 0.05$ used to classify a violation of the assumption of proportional hazards.

5.4.5 Results

$\Delta = 0.8$ and $\Delta = 1.25$ had similar impact on the relative performance of tests between case-cohort and full cohort. Hence, for clarity and brevity, only $\Delta = 1.25$ is presented. Table 5.1 shows Type 1 error rate and power for the global Cox interaction test and the global Grambsch & Therneau test. Table 5.2 and 5.3 show Type 1 error rate and power, respectively for $\Delta = 1.25$ for the single-parameter Cox interaction tests, the single-parameter Grambsch & Therneau tests, and the Pearson correlation of Schoenfeld residuals with time.

5.4.5.1 Global Tests

In the full cohort, Type 1 error for the global Cox method and global Grambsch & Therneau tests are broadly similar. Type 1 error is slightly higher in Staggered Entry than Fixed Entry, but does note exceed 5.9% for Cox Interaction and 6.6% for Grambsch & Therneau. Power is also very similar. Power increases as full cohort size and number of cases increases. Power is somewhat higher in fixed entry than staggered entry, with the largest difference in power for subcohort size 1000, sampling fraction 3% and non-case to case ratio 4:1, where power is $\sim$ 50% and 69% for staggered entry and fixed entry, respectively. This corresponds to a full cohort size of 33,333 and 248 cases.

In the case-cohort, the tests are more different. For the Cox method, Type 1 error for both weighting methods is similar, differing by at most 2%. In fixed entry, Type 1 error ranges from 4% to 6% at subcohort size 1000. At subcohort size 200, Type 1 error is similar when non-case to case ratio is 1:1, but is $\sim$ 15% when non-case to case ratio is 4:1, indicating that this may be due to the small number of cases ($\sim$ 50) seen in this scenario. In staggered entry, Type 1 error ranges from from 7% to 8% at subcohort size 1000. At subcohort size 200, Type 1 error ranges from 11% to 20%, with Type 1 error increasing as number of cases decreases. It appears, therefore, that inappropriately high Type 1 error rates are associated with smaller numbers of cases, with this exacerbated by the smaller risk set sizes seen in staggered entry.

These results are reflected in those for power of the global Cox method, where, for both entry types, at subcohort size 200 and non-case to case ratio 4:1, case-cohort power is greater than that seen in the full cohort. Under fixed entry, IPW and Prentice display similar power to each-other and to the full cohort at subcohort size 1000. Under staggered entry at subcohort size 1000, there is a greater loss of power from the full cohort than in fixed entry, and IPW displays greater power than Prentice, with this difference in power between weighting methods increasing as sampling fraction increases and non-case to case ratio decreases. The largest difference in power is seen at sampling fraction 15% and non-case to case ratio 1:1 where power is 98%, 77%, and 60% in the full cohort, IPW-weighted case-cohort and Prentice-weighted case-cohort, respectively.

Results for the global Grambsch & Therneau test are more sensitive to entry type. Under fixed entry, IPW shows somewhat higher Type 1 error than the full cohort, but still reasonable close to a nominal 5%, ranging from 5.1% to 6.9%. Type 1 error for Prentice is similar, except at non-case to case ratio 1:1 and sampling frac-

tion 15% where it is $\sim$ 10%. Power is broadly similar for both weighting systems. At subcohort size 1000, power for case-cohort methods is similar to the full cohort. At subcohort size 200, power for case-cohort methods slightly exceeds the full cohort.

Under staggered entry, both case cohort weighting methods display inappropriately high Type 1 error for Grambsch & Therneau in all scenarios. At subcohort size 200, IPW has higher Type 1 error rate than Prentice. These results are reflected in those for power, where both weighting methods tend to show greater power than the full cohort, to a greater degree for IPW at subcohort size 200.

Table 5.1: Type 1 Error and Power for Global Tests

Entry Type	N^{SC}	$\frac{N^{NC}}{N^C}$	α	Type 1 Error						Power					
				Cox Interaction			Gra. & The.			Cox Interaction			Gra. & The.		
				FC	IC	P	FC	IC	P	FC	IC	P	FC	IC	P
Fix.	200	4	3	2.9	14.7	16.5	4.8	6.6	6.0	15.6	24.5	27.2	19.5	22.7	20.2
			15	2.8	15.3	16.3	5.5	5.8	5.9	11.1	23.3	23.6	13.5	17.1	15.7
		1	3	3.5	3.8	4.8	4.0	5.0	6.3	47.7	39.4	40.1	48.2	53.9	50.8
			15	4.4	6.0	6.3	4.5	6.9	9.2	37.3	37.0	29.8	37.8	44.3	40.8
	1000	4	3	5.5	6.1	5.9	5.9	5.6	5.6	68.8	66.6	66.9	69.3	70.2	70.3
			15	4.1	5.8	6.3	4.5	5.1	4.9	49.0	47.0	45.4	49.6	49.9	49.0
		1	3	4.9	4.6	4.4	4.9	5.9	5.9	99.7	99.3	99.1	99.7	99.7	99.8
			15	4.8	5.2	5.3	4.8	5.9	10.1	97.7	96.1	92.5	97.7	97.8	97.2
Stag.	200	4	3	4.1	17.7	19.8	6.6	27.6	16.1	9.8	27.7	27.5	12.6	38.7	22.9
			15	3.5	15.5	17.4	4.7	15.3	10.7	9.7	24.8	26.5	12.2	25.2	20.8
		1	3	5.2	15.8	14.1	5.8	50.8	41.0	36.3	34.3	30.3	35.8	67.1	56.0
			15	4.8	11.9	10.7	4.7	26.3	29.3	32.9	33.8	28.9	32.9	52.9	50.8
	1000	4	3	4.9	6.6	7.8	4.9	19.2	17.4	50.0	32.6	31.4	49.5	59.1	52.6
			15	4.5	7.1	6.8	4.7	14.1	13.6	44.6	37.7	31.6	44.0	54.6	50.5
		1	3	5.9	7.7	7.6	5.7	48.0	48.1	97.7	58.5	47.7	97.8	91.9	90.1
			15	5.1	8.2	7.5	5.3	29.9	39.0	98.1	77.4	59.8	97.8	95.3	91.1

FC =Full Cohort, IC= IPW Classic, P = Prentice

N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio; α sampling fraction (%)

5.4.5.2 Single-Parameter Tests

In the full cohort, Type 1 error for the single parameter Cox method, Grambsch & Therneau, and Schoenfeld correlation tests ranges from 3.0% to 5.9%, 3.8% to 7.3%; and 4.1% to 10.6%; respectively. For the Cox method and Grambsch & Therneau

tests, entry type, full cohort size, number of cases, and correlation with the NPH covariate does not appear to have major impact on Type 1 error. Correlation with the NPH covariate does impact Schoenfeld correlation tests, with Type 1 error 4.1% to 7.0% for the uncorrelated covariate, and 4.9% to 10.6% for the correlated covariate. Power for single-parameter tests is similar, with power for all tests within 3% of each-other. Power is greater for the Normal covariate, fixed entry, larger full cohorts and more cases, with number of cases having a greater effect than full cohort size.

Again, in the case-cohort, the tests are more dissimilar. They are also more effected by entry type. In the Cox method, Type 1 error is broadly similar in both weighting methods, and does not appear to be affected by correlation with the NPH covariate except at subcohort size 200 when non-case to case ratio is 4:1, where Type 1 error rate is $\sim 11\%$ for the correlated covariate as compared to $\sim 6\%$ and $\sim 7.5\%$ for the uncorrelated covariate under fixed and staggered entry, respectively. Excluding this scenario, Type 1 error rate under fixed entry ranges from 4.4% to 6.1%. Type 1 error is higher in staggered entry than in fixed entry, ranging from 5.2% to 9.4% (with the same exclusions as in fixed entry), with Type 1 error higher at the smaller subcohort size.

Reflecting the results for Type 1 error, at subcohort size 200 when non-case to case ratio is 4:1, power of the Cox method for case-cohort methods tends to exceed that of the full cohort, to a greater degree for the binary covariate than the Normal covariate. At subcohort size 200 when non-case to case ratio is 1:1, power is close to that of the full cohort except in staggered entry for the normal covariate, where loss of power is $\sim 15\%$ for IPW and $\sim 20\%$ for Prentice. Power is within 2% of the full cohort at subcohort size 1000.

In fixed entry, Type 1 error for both weighting methods is slightly elevated from the full cohort for Grambsch % Therneau and is similar to the full cohort for correlation of Schoenfeld residuals with time. Under staggered entry, inappropriate Type 1 error in excess of that of the full cohort is seen. IPW displays noticeably higher Type 1 error than Prentice at the 3% sampling fraction and non-case to case ratio 4:1, while Prentice displays higher Type 1 error than IPW at the 15% sampling fraction and non-case to case ratio 1:1. The Grambsch & Therneau single-parameter tests display somewhat higher Type 1 error rates than those of the correlations of Schoenfeld residuals, most noticeably at the 3% sampling fraction.

Power reflects these results. In fixed entry, case-cohort methods have power within 6.5% of the full cohort. In staggered entry, power tends to exceed the full cohort.

Table 5.2: Type 1 Error for Single-Parameter Tests in NPH Datasets

Entry Type	N^{SC}	$\frac{N^{NC}}{NC}$	α	Cov	Cox Interaction			Gra. & The.			Scho. Corr.		
					FC	IC	P	FC	IC	P	FC	IC	P
Fix.	200	4	3	Corr	3.3	10.6	10.7	7.3	7.2	7.0	5.4	5.3	5.4
				UnCorr	4.0	5.7	6.1	5.0	5.4	5.2	5.0	5.0	5.1
			15	Corr	3.0	10.1	10.3	5.5	6.2	6.0	5.5	5.6	5.6
				UnCorr	3.8	5.8	5.6	5.1	5.4	5.2	4.6	4.8	4.8
		1	3	Corr	3.9	5.0	4.7	4.5	6.0	5.6	6.4	6.9	6.6
				UnCorr	4.7	4.9	5.2	5.0	6.2	6.5	4.7	4.7	5.6
			15	Corr	4.6	5.8	5.6	4.8	6.5	6.2	5.3	5.8	6.2
				UnCorr	4.5	5.4	5.2	4.5	6.4	7.1	4.9	5.2	5.6
	1000	4	3	Corr	4.9	5.5	5.4	5.4	5.9	6.1	7.3	7.2	7.4
				UnCorr	4.0	4.5	4.4	4.4	4.5	4.5	4.4	4.3	4.4
			15	Corr	5.2	5.1	5.4	5.4	5.3	5.5	7.3	7.3	6.9
				UnCorr	5.0	5.3	5.2	5.0	5.4	5.3	5.4	5.5	5.7
		1	3	Corr	3.7	3.9	3.8	3.8	4.9	5.1	10.6	10.8	10.5
				UnCorr	5.5	5.8	5.6	5.6	6.9	7.1	5.4	5.6	5.5
			15	Corr	4.9	5.6	5.1	5.1	6.5	6.4	7.2	7.3	8.2
				UnCorr	5.7	4.8	5.0	5.6	5.4	6.2	5.0	5.8	6.2
Stag.	200	4	3	Corr	3.5	10.6	12.0	5.1	13.1	8.5	4.9	9.5	7.5
				UnCorr	3.8	7.5	7.6	5.0	15.7	8.4	4.9	10.4	7.2
			15	Corr	5.0	12.5	14.2	6.1	11.9	9.5	5.1	9.1	8.2
				UnCorr	3.9	6.8	7.1	4.7	10.1	7.3	4.7	7.1	6.2
		1	3	Corr	5.7	8.6	8.7	6.2	22.0	18.4	6.4	18.0	14.4
				UnCorr	3.7	9.1	9.4	3.9	25.6	21.0	4.1	21.6	19.6
			15	Corr	5.9	8.5	9.4	6.5	14.3	15.3	6.8	14.1	14.8
				UnCorr	5.6	8.9	9.3	5.2	16.8	19.3	5.0	15.7	16.6
	1000	4	3	Corr	5.9	6.6	6.8	6.1	10.6	9.2	6.3	9.9	8.9
				UnCorr	5.2	6.8	6.4	5.5	15.4	13.7	5.3	13.9	12.8
			15	Corr	5.9	6.0	6.1	6.0	8.5	8.7	7.7	10.3	10.2
				UnCorr	4.9	6.8	6.5	5.2	10.3	9.7	5.7	9.0	8.9
		1	3	Corr	5.4	6.6	6.6	5.5	19.3	19.7	7.4	20.8	19.6
				UnCorr	5.9	7.7	6.6	5.5	26.1	26.5	5.4	24.9	25.4
			15	Corr	4.9	5.6	6.4	5.1	15.1	18.4	5.8	14.2	16.0
				UnCorr	5.5	5.2	6.2	5.5	16.4	20.6	7.0	15.4	20.0

FC =Full Cohort, IC= IPW Classic, P = Prentice

N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio; α sampling fraction (%)

Table 5.3: Power for Single-Parameter Tests in NPH Datasets

Entry Type	N^{SC}	$\frac{N^{NC}}{N^{C}}$	α	Cov	Cox Interaction			Gra. & The.			Scho. Corr.		
					FC	IC	P	FC	IC	P	FC	IC	P
Fix.	200	4	3	Bin	6.2	10.3	10.7	8.6	10.1	9.8	9.1	9.2	9.1
				Norm	26.4	24.4	27.1	26.5	30.5	28.9	25.5	26.0	26.3
			15	Bin	5.8	11.1	11.2	7.7	7.9	7.9	7.0	6.7	6.9
				Norm	18.1	20.7	21.0	18.4	21.3	20.0	17.8	19.2	18.8
		1	3	Bin	13.3	12.9	13.3	13.6	16.5	16.3	14.7	15.0	15.5
				Norm	62.3	58.2	55.1	62.4	68.6	62.1	61.9	64.3	60.7
			15	Bin	11.1	12.3	11.7	11.2	13.8	13.2	11.5	11.9	11.8
				Norm	53.3	50.6	42.1	53.4	57.9	51.7	52.1	55.1	50.3
	1000	4	3	Bin	18.9	17.6	17.6	19.6	19.5	19.5	20.6	20.6	20.7
				Norm	83.1	81.3	80.5	83.1	83.6	82.7	82.9	83.1	82.1
			15	Bin	12.9	13.1	13.3	13.3	13.8	14.0	13.1	13.1	13.7
				Norm	62.7	60.5	59.3	62.9	63.3	62.9	62.1	62.7	61.1
		1	3	Bin	42.5	39.2	38.2	42.4	42.6	41.5	45.1	45.3	44.5
				Norm	100.0	99.9	99.9	100.0	99.9	99.9	100.0	99.9	99.9
			15	Bin	36.5	34.8	34.1	35.8	36.8	37.3	33.0	33.8	33.5
				Norm	99.2	98.3	97.0	99.3	98.7	98.4	98.9	98.6	98.7
Stag.	200	4	3	Bin	4.9	13.0	13.0	6.6	14.5	9.4	6.3	11.2	8.5
				Norm	19.2	25.2	23.8	19.5	36.6	23.8	18.5	28.3	23.4
			15	Bin	5.3	12.3	14.0	8.1	12.4	9.7	6.9	9.4	8.3
				Norm	17.3	19.5	19.7	15.9	24.2	20.8	15.4	20.4	18.9
		1	3	Bin	11.4	13.0	12.5	11.5	26.9	22.5	11.6	24.4	20.3
				Norm	53.9	36.1	33.4	53.9	57.8	49.3	52.3	51.9	46.9
			15	Bin	11.1	12.7	12.5	11.6	21.9	20.1	10.2	18.7	17.8
				Norm	48.1	34.9	32.0	45.8	48.0	45.0	44.9	44.3	43.4
	1000	4	3	Bin	12.3	12.2	11.7	12.7	17.9	16.1	13.8	16.8	15.4
				Norm	62.9	44.3	41.3	62.8	65.2	58.0	62.9	63.1	58.1
			15	Bin	11.6	13.2	12.8	11.3	17.4	16.4	13.1	16.1	15.4
				Norm	60.7	48.5	44.2	59.8	60.1	55.8	59.6	58.1	55.4
		1	3	Bin	29.6	19.5	17.4	29.5	36.2	33.1	30.9	37.1	32.9
				Norm	99.5	68.2	59.9	99.5	91.0	88.0	99.5	90.8	88.3
			15	Bin	33.3	23.3	20.6	32.1	38.2	36.6	25.6	32.9	29.6
				Norm	98.9	87.2	74.1	98.7	94.9	91.0	98.6	94.4	90.1

FC =Full Cohort, IC= IPW Classic, P = Prentice

N^{SC} Subcohort size; $\frac{N^{NC}}{N^{C}}$ non-case to case ratio; α sampling fraction (%)

5.5 Discussion

The results from this chapter indicate that where risk sets at individual failure times are small or more variable, tests based on correlation of Schoenfeld residuals with $g(t)$ are inappropriate in the case-cohort design. Further, the scaled Schoenfeld residuals tests of Grambsch & Therneau as implemented in STATA are also inappropriate. This is likely due to the reasoning given by Winnett and Sasieni (2001), that the substitution of the average variance for the weighted variance of the covariates at each failure time may be inappropriate when variance of covariates changes substantially over failure times. It is possible that use of the time-specific weighted variances rather than the average in the case-cohort would result in improvements in performance by taking this into account. While the individual variances at each failure time are readily obtainable in R, they are not recorded in STATA, and hence this approach was not further investigated in this thesis.

Cox's method is more promising as a method for detection of non-proportional hazards in case-cohort samples. Single-parameter tests following global inclusion of interactions displayed improved Type 1 error rate and less loss of power than a global test of all interactions, and have the advantage of facilitating identification of the specific covariates(s) that display non-proportional hazards, where present, while still accounting for correlation between covariates with proportional hazards and those with non-proportional hazards. IPW weighting showed a small improvement in Type 1 error rate and power over Prentice weighting in some scenarios with this test. As such, global inclusion of interactions in an IPW-weighted Cox model, followed by single-parameter Wald tests is recommended for detection of non-proportional hazards in the case-cohort. Where risk set sizes are small or highly variable, number of cases is small, covariates are continuous, or covariates are highly correlated, results indicating presence of non-proportional hazards should be regarded with caution.

The simulation study presented in this chapter displays an example of where results from simulations analysed under fixed entry may not apply to datasets analysed under staggered entry. While the impact of time-scale was small in Chapter 3, where estimates can be considered to be aggregated over risk sets, they were more impactful here, where values from individual risk sets had more influence. The findings of Xue et al. (2013) regarding Prentice-weighted Schoenfeld residuals are not contradicted by the results presented in this Chapter, rather, a wider range of risk set sizes and risk set variation is considered here.

The simulation study did not consider a wide range of forms of non-proportional hazards, nor did it consider a wide range of ϕ or β. It is possible that varying forms of non-proportional hazards or values of ϕ and β may have differing impacts on the case-cohort sample. However, in this simulation study, there was very little difference seen in relative performance of full cohort and case-cohort methods between the two values of Δ considered, 0.8 and 1.25. Further, Xue et al. (2013) also found only small differences in performance of Prentice-weighted Schoenfeld residuals for a wider range of forms of non-proportional hazards.

Chapter 6

Model & Variable Selection

6.1 Introduction

Methods for model selection in maximum likelihood estimation include Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and, for nested models, Likelihood ratio (LR) and Wald tests. These methods can also be used for variable selection. In the proportional hazards model, for the full cohort, large-sample properties and tests based on the likelihood ratio method and an asymptotic chi-squared distribution are justified in the case where there is a partial likelihood rather than a likelihood, under broad conditions (Cox, 1975). Wald tests are valid in the case-cohort design when appropriate methods have been used for variance estimation, but there has been little work on the other methods in the case-cohort design. The field of complex survey sampling, however, has seen work which may apply to the case-cohort model. Lumley and Scott (2013) considered the case-cohort design as a special case of complex sampling when they introduced a modified Likelihood ratio test dLR. For complex survey sampling, though not discussing the case-cohort design, Xu et al. (2013) proposed replacement of the likelihood with the pseudolikelihood for a modified BIC ($pBIC$), and Lumley and Scott (2015) built on their 2013 paper to describe modifications to AIC and BIC ($dAIC$ and $dBIC$).

In this chapter I first, in Sections 6.2 and 6.3, describe the methods mentioned above. In Section 6.4 I perform a simulation study comparing naïeve replacement of the likelihood with the pseudolikelihood (denoted $pAIC$, $pBIC$, and pLR) with the modifications of Lumley & Scott ($dAIC$, $dBIC$, and dLR) in the case-cohort, and with the standard methods in the full cohort. The effects of sparse data are also demonstrated. In the absence of sparse data, dLR is found to have similar power to robust Wald tests, with Type 1 error rate approximately 5%. In the presence of sparse data, dLR is superior to robust Wald tests. In the absence of sparse data,

$dBIC$ shows little difference from the naieve use of the pseudo-log-likelihood in the standard BIC formula ($pBIC$). In the presence of sparse data $dBIC$ shows reduced power to select the true model, and $pBIC$ is superior. $dAIC$ shows improvement in power to select the true model over naieve methods. Where subcohort size and number of cases is not overly small, loss of power from the full cohort for $dAIC$, $dBIC$ and $pBIC$ is not substantial.

6.2 Wald Test and Likelihood Ratio Test

Full cohort methods for assessing significance of variables in a model include the Likelihood ratio test, the Wald test, and the score Test. These tests are asymptotically equivalent, but make different approximations in small samples. The score test assesses improvement based on movement towards the alternative from the null, the Wald test assesses improvement based on movement to the null from the alternative, and the Likelihood ratio test directly compares the two hypotheses (Engle, 1984).

Define a full model M_M with k_M parameters β_M, and likelihood L_M. Consider the partition of β_M into β_1 of dimension k_1 and β_2 of dimension k_2. Define submodel M_1 with k_1 parameters β_1 and submodel M_2 with k_2 parameters β_2. Consider an analysis where the aim is to assess whether the covariates corresponding to β_2 contribute to the model M_M, perhaps in order to choose between models M_M and M_1.

The Wald test (Wald, 1943) assesses whether all elements of β_2 are simultaneously equal to 0. A Taylor series approximation to the score function about the MLE finds that estimator $\hat{\beta}$ is approximately normal with mean β and variance V equal to the inverse of the Fisher Information I. Let V_M be the variance of $\hat{\beta}_M$ and let V_2 refer to the submatrix of V_M corresponding to β_2. The Wald test statistic is defined as:

$$W = \frac{\hat{\beta}_2^2}{\hat{V}_2} \text{ for } k_2 = 1$$

$$W = \hat{\beta}_2'(\hat{V}_2)^{-1}\hat{\beta}_2 \text{ for } k_2 > 1$$

The Likelihood ratio test statistic (Neyman and Pearson, 1928), can be defined:

$$LR = -2ln\frac{L_1}{L_M} = 2(ln(L_M) - ln(L_1))$$

Both test statistics are asymptotically chi square with degrees of freedom k_2. While the Likelihood Ratio Test requires fitting of both M_M and M_1, the Wald Test only requires fitting of M_M. However, note that the Wald test of a parameter is valid

only if the profile likelihood for the parameter is well approximated by a normal likelihood (Pawitan, 2000).

6.2.1 Case-Cohort Implementation

The Wald test relies only on estimates of β_2 and V_2, given by M_M. As such, it is theoretically justified in the case-cohort model, despite the presence of a pseudopartial likelihood rather than a partial likelihood, as long as the variance estimates of β_2 are appropriate for the case-cohort design. The Likelihood ratio test, by contrast, relies upon the values of the likelihoods for M_M and M_1. A naïeve modification for the case-cohort model would be to replace the likelihoods in the formula with the pseudopartial likelihoods pL^* from the case-cohort models to give pLR, and assume that pLR is also asymptotically chi square with degrees of freedom k_2, such that

$$pLR = -2ln\frac{pL_1^*}{pL_M^*} = 2(ln(pL_M^*) - ln(pL_1^*))$$

Lumley and Scott (2013) introduced a modified Likelihood Ratio test, dLR, with identical test statistic to the naïeve pseudopartial likelihood substitution pLR. The modification is in the distribution of the test statistic. In complex sampling, under the null hypothesis that $\beta_2 = 0$, the test statistic converges in distribution to Q_2, a linear combination of k_2 independent χ_1^2 random variables. The coefficients (de_i) of this linear combination are the eigenvalues of the design effects matrix DE. Let $\hat{V}^n$ refer to the na"ieve variance estimate and let $\hat{V}$ refer to the design-based variance estimate, accounting for the case-cohort design. Let the observed information matrix $\hat{\mathcal{I}}$ be the inverse of the naieve variance estimate matrix $\hat{V}^n$. As before, let subscript M refer to the maximal model. Let $\hat{V}_{Mij}$ refer to the submatrix of $\hat{V}_M$ corresponding to the covariance of $\hat{\beta}_i$ with $\hat{\beta}_j$ and let $\hat{\mathcal{I}}_{Mij}$ refer to the corresponding submatrix of $\hat{\mathcal{I}}_M$.

DE is estimated from the full model M_M, with

$$\hat{DE} = (\hat{\mathcal{I}}_{M22} - \hat{\mathcal{I}}_{M21}\hat{\mathcal{I}}_{M11}^{-1}\hat{\mathcal{I}}_{M12})\hat{V}_{M22}$$

The distribution of Q_2 can be approximated by matching the moments, via their cumulants q, to a known distribution. One method for this approximation is that of Satterthwaite (1946). This approximation equates the first two moments of Q_2 with those of a $\Gamma(\hat{g}, \hat{\theta})$ distribution, with a final test distribution of $\Gamma(\hat{g}, \frac{dLR}{\hat{\theta}})$, where:

$$q_1 = \sum_{i=1}^{k_2} de_i, \qquad q_2 = 2\sum_{i=1}^{k_2}(de_i)^2, \qquad \hat{g} = \frac{q_1^2}{q_2}, \qquad \hat{\theta} = \frac{q_2}{q_1}$$

.

6.3 AIC and BIC

For a model fitting k parameters β with likelihood L, the Akaike Information Criterion (Akaike, 1974) for the model can be defined as:

$$AIC = -2ln(L) + 2k$$

and where n is the sample size, the Bayesian Information Criterion (Schwarz, 1978)

$$BIC = -2ln(L) + k \times ln(n)$$

When choosing between models, the candidate model with the smallest criterion value is preferred. The criteria differ in the penalty term used to penalise for inclusion of additional explanatory variables. Use of BIC requires a choice of n; Volinsky and Raftery (2000) proposed that in the Cox model, n be defined as the number of failures, rather than the number of individuals or observations, with the justification that this corresponds to a more realistic prior on the parameter space.

6.3.1 Case-Cohort Implementation

As with the Likelihood Ratio test, in the case-cohort design, a basic adaptation for both methods would be to use the pseudopartial likelihood pL^* as a substitute for the partial likelihood from the full cohort, such that:

$$pAIC = -2ln(pL^*) + 2k$$

$$pBIC = -2ln(pL^*) + k \times ln(n)$$

For variable and model selection in complex survey data, Xu et al. (2013) propose this adaptation of BIC for survey data to give $pBIC$, and show that, if one or more of the models is true, then the probability that the most parsimonious true model is selected converges to one as $n \to \infty$. The case cohort design is not explicitly mentioned.

Lumley and Scott (2015) used similar theory to their work on the Likelihood Ratio test to propose principled survey analogues of AIC and BIC, $dAIC$ and $dBIC$, where the second term penalises for larger design effects as well as for increasing numbers of parameters. The case-cohort design is not mentioned, but, following the same logic as in their 2013 paper, can be considered as a special case of such models.

6.3.1.1 dAIC

For $dAIC$ criterion, the pseudopartial likelihood is substituted for the likelihood in the AIC formula, and the penalty term, $-2k$ in AIC, is scaled by a quantity $\bar{\delta}\epsilon$. $\bar{\delta}\epsilon$ is defined as k^{-1} times the trace of the design effects matrix ΔE, where ΔE is estimated as the product of the observed information matrix $\hat{\mathcal{I}}$ and the design-based variance estimate $\hat{V}$.

$$\hat{\bar{\delta}}\epsilon = k^{-1}tr(\hat{\mathcal{I}}\hat{V})$$

$$dAIC = -2ln(pL^*) + 2k\hat{\bar{\delta}}\epsilon$$

$dAIC$ will place a greater penalty on additional parameters than $pAIC$ where estimates of design-based variance are larger than those of naïeve variance.

Note that ΔE, the design effects matrix for $dAIC$ is distinct from DE, the design effects matrix for dLR. When comparing models, ΔE is estimated separately for each candidate model without reference to any other models. DE, by contrast is estimated for submodel M_1 with reference to a maximal model M_M.

6.3.1.2 dBIC

For a modified BIC for complex survey sampling, Lumley & Scott conceptualize candidate models as submodels of a maximal model M_M fitted on k_M parameters β_M. A particular submodel, fitted on k_1 parameters β_1 can hence be conceptualised as setting the remaining $k_2 = k_M - k_1$ parameters β_2 equal to 0. The approach is similar to the Wald test for nested models, but unlike the Wald test, the submodels of M_M need not be nested within *eachother* to be directly compared. Lumley & Scott describe the substitution of W_D the "design-based" Wald statistic for $\beta_2 = 0$, for the first term in the BIC formula. Note that this Wald statistic is that derived from the naïeve estimate of the variance-covariance matrix, and so their terminology is different in that respect from that used in this thesis. The penalty term is $k_2 ln(n/\hat{\bar{d}}e)$, where $\bar{d}e$ is the geometric mean of the eigenvalues of the design effect matrix DE.

$$dBIC = W_D - (k_2)ln(n/\hat{\bar{d}}e)$$

For calculation of $dBIC$ for M_1, a submodel of M_M, the design effect matrix is the same as that which would be used in the calculation of the distribution of the test statistic of dLR for comparison of the models M_1 and M_M. Hence, when using $dBIC$ to compare multiple submodels of a maximal model M_M, a unique DE and W_D are calculated for each candidate model, but for each submodel they are calculated with reference to the same maximal model M_M. The submodel with the smallest $dBIC$ is preferred. Note that under this formulation, $dBIC$ for the maximal model is 0.

Lumley & Scott noted that if design effects are large, $pBIC$ will overestimate the amount of information in the sample and choose a more complex model, but that if design effects are less than one, as can happen with an effective stratification, $pBIC$ will underestimate the amount of information and prefer simpler models. They concluded that while $dAIC$ and $dBIC$ may often select identical models to $pAIC$ and $pBIC$, $dAIC$ and $dBIC$ will be more accurate in some circumstances, and so should be preferred. However, the Wald test statistic included in the $dBIC$ formula, has potential to be invalid where the profile pseudolikelihood is non-normal.

6.4 Simulation Study

The purpose of this simulation study is to compare the performance of the above methods in the full cohort and the case cohort, in the presence and absence of sparse data. To my knowledge, comparison of dLR and Wald tests in the case-cohort with sparse data has not previously been investigated, nor has use of $pBIC$, $dBIC$, $pAIC$ and $dAIC$ in the case-cohort design.

6.4.1 Data Generating Mechanism

Data generation was as described in Chapter 2, with the following specifications:

Four binary covariates, $X1$, $X2$, $X3$ and $X4$ were generated from Binomial(1, 0.5). $X1$ and $X2$ were correlated with $\rho = 0.5$, and $X3$ and $X4$ were generated independently. Survival times were generated from $X1$ and $X3$, each with equal coefficients of ln(HR) per standard deviation of the covariate, for Hazard Ratios 1.1 and 1.3.

To demonstrate the effect of sparse data, a similar simulation study was performed for subcohort size 1000, sampling fraction 3% and non-case to case ratio 1:1, with the data generated so as to produce sparse data as might yield a non-normal profile likelihood (Greenland (1986), Cole et al. (2014)).

$X1$, $X2$ and $X4$ were generated as above. $X3$ was generated from Binomial(1, .01). Survival times were generated from $X1$ and $X3$. β for $X1$ was equal to ln(1.1)/SD. Two values of β for $X3$ were considered; ln(0.92)/SD and ln(0.85)/SD. Only full cohorts where precisely 2 cases had $X3 = 1$ were accepted, with full cohorts not meeting this condition redrawn.

6.4.2 Targets

6.4.2.1 Hypothesis Testing

Wald and Likelihood ratio tests were assessed as follows. In the notation of Section 6.2, the maximal model M_M contained 4 covariates, and featured $2^4 - 1 = 15$ potential submodels M_{1i} each with k_{1i} estimated parameters $\hat{\beta}_{1i}$. Test parameters were defined as the remaining k_{2i} parameters β_{2i} not included in M_{1i}. For example, let γq be the parameter for covariate Xq. Then, for the submodel containing only $X1$, $X2$ and $X3$, the test parameter is $\gamma 4$, corresponding to $X4$. The null hypothesis for each set of test parameters β_{2i} is that in the maximal model M_M, $\beta_{2i} = 0$. The target is the evaluation of the null hypothesis. Hence when $\gamma 1$, $\gamma 3$, or both $\gamma 1$ and $\gamma 3$ were included in the test parameters, the target was rejection of the null hypothesis. When neither $\gamma 1$ nor $\gamma 3$ were included in the test parameters, the target was failure to reject the null hypothesis. While Wald and Likelihood ratio methods could be used to select a preferred model by means of e.g. backward selection, this target was not considered.

6.4.2.2 Model Selection

$pAIC$, $pBIC$, $dAIC$ and $dBIC$ were assessed as follows. Including the null model, the maximal model, and all its submodels, there were $2^4 = 16$ candidate models. The target was selection of the true model, the model containing only $X1$ and $X3$.

6.4.3 Methods

Full cohort, IPW Classic and Prentice-weighted estimation methods were considered. The "robust" Huber sandwich estimator was used for estimation of the design-based variance.

6.4.3.1 Hypothesis Testing

For each set of test parameters, test-statistics with reference to the maximal model M_M were calculated, together with their associated p-values. A cutoff criteria of $p < 0.05$ was used to reject the null hypothesis. The Satterthwaite (1946) method was used to approximate the distribution of dLR.

6.4.3.2 Model Selection

$pAIC$, $pBIC$, $dAIC$ and $dBIC$ were calculated for each of the candidate models, and for each criterion, the model with the smallest value was selected. Following the recommendation of Volinsky and Raftery (2000). N^C was used for the value of n for BIC, $dBIC$ and $pBIC$.

6.4.4 Performance Measures

6.4.4.1 Hypothesis Testing

For each set of test parameters, power was calculated as the proportion of replicates where the null hypothesis was rejected when test parameters contained at least one of $\gamma 1$ and $\gamma 3$, and Type 1 error rate was calculated as the proportion of replicates where the null hypothesis was rejected for each set of test parameters that did not contain at least one of $\gamma 1$ and $\gamma 3$.

6.4.4.2 Model Selection

The proportion of replicates where the true model was selected was calculated. As a secondary measure of interest, to illustrate the differing behaviours of the methods in particular circumstances, the proportion of replicates where each submodel model was selected was also calculated.

6.4.5 Results

In this section I first describe the results for hypothesis testing and model selection in the data-generating mechanism that was not designed to create sparse data, followed by presenting results in the presence of sparse data.

In both the full cohort and the case-cohort, results for all methods are most influenced by β and N^C, the number of cases in the dataset. Full Cohort size and number of controls per case in the corresponding case-cohort sample have minimal influence. Entry type and case-cohort weighting system also had only minor effects. Hence, for clarity and brevity, presented here are results for staggered entry, IPW Classic weighting, and the 3% sampling fraction. The full cohort size and number of cases corresponding to each combination of subcohort size, non-case to case ratio, and sampling fraction was shown in Chapter 2 and is reproduced here for reference.

Table 6.1: Full Cohort Sizes and Number of Cases Considered in this Thesis

N^{SC} : Subcohort size		200		1000	
$\frac{N^{NC}}{N^C}$ Non-case to case ratio	α (%) Sampling Fraction	N Full cohort size	N^C No. of cases	N Full cohort size	N^C No. of cases
4	15	1,333	48	6,667	241
	3	6,667	50	33,333	248
1	15	1.333	170	6,667	850
	3	6,667	194	33,333	970

6.4.5.1 Hypothesis Testing

Table 6.2 shows results for Type 1 error in the full cohort and case cohort. *pLR* showed inappropriate Type 1 error rate ranging from 6.4% to 26.1% (MCSE 0.8% - 1.4%), with Type 1 error rate higher at lower numbers of controls per case. Type 1 error rate and associated MCSE was broadly similar for the remaining tests, ranging from 2.8% to 6.6% for Type 1 error rate and 0.5% to 0.8% for MCSE. MCSE bounds for these tests encompassed a nominal alpha level of 5% in all scenarios except for *LR* in the full cohort at subcohort size 200, HR/SD 1.1, $\frac{N^{*NC}}{N^C}$=1 and test parameters $\gamma 2$ $\gamma 4$, where it was slightly elevated (5.1%-8.1%), and for both Wald and *dLR* in the case cohort for subcohort size 1000, where at HR/SD 1.1, $\frac{N^{*NC}}{N^C}$=4, and test parameter $\gamma 4$, it was slightly depressed (3.3%-4.9% for Wald and 2.6%-5% for *dLR*); and at HR/SD 1.3, $\frac{N^{*NC}}{N^C}$-1. and test parameters $\gamma 2$ $\gamma 4$ it was somewhat depressed (1.9%-3.9% for Wald and 1.8%-3.8% for *dLR*).

Table 6.2: Type 1 Error Rate for Hypothesis Tests of Parameters in M_M

(Staggered Entry, sampling fraction 0.03, and IPW Classic)

$N^{SC}_{(0)}$:			200					1000				
HR/SD	$\frac{N^{NC}}{N^C}$	Test Parameters	Full Cohort		Case Cohort			Full Cohort		Case Cohort		
			Wald	LR	Wald	dLR	pLR	Wald	LR	Wald	dLR	pLR
1.1	4	$\gamma 2$	4.2	4.3	5.2	5.6	9.5	5.0	5.2	5.5	5.5	8.6
		$\gamma 4$	3.9	4.3	5.3	5.5	9.2	4.4	4.6	3.7	3.8	6.4
		$\gamma 2$ $\gamma 4$	4.2	4.7	4.7	4.6	12.0	4.2	4.3	4.3	4.1	8.7
	1	$\gamma 2$	5.2	5.3	4.3	4.5	20.4	5.3	5.3	4.9	4.9	17.0
		$\gamma 4$	5.7	5.9	5.3	5.3	18.5	5.0	5.1	6.3	6.3	18.4
		$\gamma 2$ $\gamma 4$	6.4	6.6	4.5	4.4	2.6	4.9	4.9	5.2	5.1	2.5
1.3	4	$\gamma 2$	3.9	4.1	3.9	4.2	9.3	4.4	4.4	5.0	5.2	7.7
		$\gamma 4$	4.5	5.1	5.0	5.5	10.9	4.7	4.7	5.9	6.2	8.8
		$\gamma 2$ $\gamma 4$	4.2	4.7	4.2	4.5	10.3	4.6	4.6	4.8	5.0	10.5
	1	$\gamma 2$	5.2	5.1	4.2	4.1	18.0	4.8	4.8	4.2	4.2	17.2
		$\gamma 4$	4.4	4.7	4.8	5.0	20.3	5.2	5.2	5.0	5.0	19.7
		$\gamma 2$ $\gamma 4$	4.3	4.2	5.3	5.3	26.1	4.5	4.6	2.9	.28	25.6

N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio

Table 6.3 shows full cohort and case cohort results for power. As *pLR* showed highly inappropriate Type 1 error, results for *pLR* are not presented for power. For any particular combination of "true" test parameters $\gamma 1$ and $\gamma 3$ included in the test parameters, results were similar when either $\gamma 2$, $\gamma 4$ or both $\gamma 2$ and $\gamma 4$ ("false" test parameters) were also included. Hence, for clarity and brevity the mean of power

for such combinations is presented. Wald tests and dLR in the case-cohort show less power than their equivalents in the full cohort. Loss of power from the full cohort increases with subcohort size and number of cases in the case-cohort sample, and is not dissimilar from that seen in Fig. A.1b indicating that loss of power is due to the lesser information in the case-cohort.

Table 6.3: Power for Hypothesis Tests of Parameters included in M_M

(Staggered Entry, sampling fraction 0.03, and IPW Classic)

$N^{SC}_{(0)}$:				200				1000			
HR/SD	$\frac{N^{NC}}{N^C}$	Test Parameters		Full Cohort		Case Cohort		Full Cohort		Case Cohort	
		true	false	Wald	LR	Wald	dLR	Wald	LR	Wald	dLR
1.1	4	$\gamma1$ $\gamma3$	none	11.9	12.9	9.4	9.9	37.6	38.3	32.7	32.7
			$\gamma2$ and/or $\gamma4$	10.0	11.8	8.4	8.8	35.3	36.0	28.1	28.1
		$\gamma1$	none	9.2	9.9	8.2	8.7	24.8	25.0	20.6	20.7
			$\gamma2$ and/or $\gamma4$	7.6	8.7	6.9	6.9	20.8	21.1	17.7	17.6
		$\gamma3$	none	10.2	10.8	9.1	9.8	31.2	31.5	25.4	25.8
			$\gamma2$ and/or $\gamma4$	7.8	8.8	7.1	6.8	22.8	23.0	17.6	18.0
	1	$\gamma1$ $\gamma3$	none	31.1	31.8	14.4	14.9	94.1	94.1	69.5	69.7
			$\gamma2$ and/or $\gamma4$	28.5	29.1	12.9	13.8	92.9	92.9	64.7	64.9
		$\gamma1$	none	18.3	18.6	9.9	9.8	71.8	71.8	43.7	43.7
			$\gamma2$ and/or $\gamma4$	15.9	16.3	8.8	9.0	67.1	67.3	38.0	38.3
		$\gamma3$	none	29.7	29.9	16.2	16.5	84.1	84.1	54.5	54.7
			$\gamma2$ and/or $\gamma4$	19.9	20.3	10.8	11.3	74	74	42	42
1.3	4	$\gamma1$ $\gamma3$	none	55.1	56.8	44.3	46	99.9	99.9	99.4	99.3
			$\gamma2$ and/or $\gamma4$	48.9	52.7	38.5	39.9	99.8	99.8	99.1	99.2
		$\gamma1$	none	35.2	35.9	27.2	28.4	95.3	95.5	87.3	87.4
			$\gamma2$ and/or $\gamma4$	29.4	31.8	22.6	23.1	94.5	94.6	86.5	86.5
		$\gamma3$	none	41.9	43.2	34.8	36.4	98.5	98.5	94.7	95.0
			$\gamma2$ and/or $\gamma4$	29.9	32.2	24.4	26.0	95.6	95.7	88.2	88.6
	1	$\gamma1$ $\gamma3$	none	98.6	98.6	83.5	83.9	100	100	100	100
			$\gamma2$ and/or $\gamma4$	98.5	98.5	81.4	82.1	100	100	100	100
		$\gamma1$	none	87.6	87.6	58.6	59.1	100	100	99.9	99.9
			$\gamma2$ and/or $\gamma4$	85.4	86.0	54.3	54.6	100	100	99.9	99.9
		$\gamma3$	none	93.8	93.9	66.8	67.2	100	100	99.9	99.9
			$\gamma2$ and/or $\gamma4$	87.6	87.8	53.9	54.2	100	100	99.9	99.9

N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio

6.4.5.2 AIC and BIC

Tables 6.5 and 6.4 show selection rates of the true model by AIC, BIC, and their case-cohort modifications. Additionally, selection rates for certain other models and combinations of models is shown to illustrate the behaviour of the methods.
In the full cohort, selection of the true model by both AIC and BIC was low where β was small and/or number of cases was low. In the case-cohort, $pBIC$ and $dBIC$ perform similarly to each-other, with $pBIC$ slightly more parsimonious. Loss of power to select the true model from the full cohort was not substantial, especially at the higher β, with both $pBIC$ and $dBIC$ in the case-cohort somewhat less parsimonious than BIC in the full cohort, as indicated by a greater tendency to select a model including an additional covariate as well as $X1 and X3$, and a lesser tendency to select the null model at the lower β.

$dAIC$ and $pAIC$ are much more dissimilar. In general, $dAIC$ gives results more similar to the full cohort than pAIC for selection of the true model. Loss of power to select the true model from the full cohort was not substantial for $dAIC$, especially at the higher β. $pAIC$ is substantially less parsimonious than $dAIC$, as indicated by a lesser tendency to select the null model at the lower β and a greater tendency to select a model including an additional covariate as well as $X1 and X3$.

As in hypothesis testing, loss of power from the full cohort for modifications of AIC and BIC decreased as subcohort size and number of cases increased. Substitution of the correlated $X2$ for the true $X1$ did not appear to be overly different in case-cohort methods than the full cohort. Selection rates for a univariate model containing $X1$ or $X3$ were similar in full cohort and case-cohort methods.

Table 6.4: Model Selection: Percentage Selection Rates for BIC and Modifications

(Staggered Entry, sampling fraction 0.03, and IPW Classic)

HR/SD:			1.1			1.3		
Model	α	$\frac{N^{*NC}}{N^C}$	BIC	pBIC	dBIC	BIC	pBIC	dBIC
True Model	200	4	**1.0**	**2.4**	**2.6**	**16.5**	**17.4**	**17.6**
	200	1	**2.1**	**3.9**	**4.9**	**76.0**	**49.0**	**48.4**
	1000	4	**3.2**	**4.1**	**4.6**	**88.8**	**79.7**	**79.4**
	1000	1	**36.5**	**29.4**	**31.1**	**98.0**	**85.8**	**82.3**
True Model And Only One of X2 or X4	200	4	0.05	0.45	0.45	0.8	1.95	2.35
	200	1	0	0.8	1.2	1.05	6.8	8.7
	1000	4	0.2	0.3	0.4	1.35	3.6	3.9
	1000	1	0.5	1.8	2.5	1.0	6.85	8.5
Univariate Model X1 or X3	200	4	8.15	9.45	9.5	22.0	18.6	18.25
	200	1	12.55	13.25	13.2	8.95	10.8	9.05
	1000	4	14.2	14.95	15.4	3.3	4.95	4.8
	1000	1	22.8	19.95	19.3	0	0	0
X3 and X2 Only	200	4	0.2	0.7	0.7	1.8	2.9	3.0
	200	1	0.8	3.1	3.9	1.9	4.7	5.1
	1000	4	0.3	1.1	1.4	1.2	2.0	2.0
	1000	1	2.0	3.8	4.4	0	0.1	0.1
Null Model	200	4	73.3	60.6	59.4	27.9	25.8	24.5
	200	1	65.5	46.2	41.0	0.7	3.0	2.4
	1000	4	63.2	57.1	54.9	0.2	0.2	0.2
	1000	1	13.5	13.9	10.9	0	0	0

N^{SC} Subcohort size; $\frac{N^{NC}}{N^C}$ non-case to case ratio

Table 6.5: Model Selection: Percentage Selection Rates for AIC and Modifications

(Staggered Entry, sampling fraction 0.03, and IPW Classic)

HR/SD:			1.1			1.3		
Model	N^{SC}	$\frac{N^{NC}}{N^{C}}$	AIC	pAIC	dAIC	AIC	pAIC	dAIC
True Model	200	4	**4.7**	**5.0**	**4.0**	**30.3**	**22.7**	**22.4**
	200	1	**13.5**	**8.8**	**5.6**	**66.2**	**35.5**	**49.9**
	1000	4	**18.7**	**16.5**	**13.8**	**70.5**	**60.5**	**69.1**
	1000	1	**60.7**	**32.6**	**37.9**	**71.2**	**42.9**	**69.8**
True Model And Only One of X2 or X4	200	4	0.9	1.9	1.0	5.05	7.6	4.15
	200	1	2.35	5.25	1.3	13.7	18.85	8.9
	1000	4	2.85	3.95	2.7	13.05	16.05	12.3
	1000	1	10.35	14.55	6.25	13.25	23.35	14.1
Univariate Model X1 or X3	200	4	12.75	12.4	12.3	15.65	14.1	17.7
	200	1	15.95	10.7	13.8	0.8	2.35	8.7
	1000	4	17.15	15.35	17.25	0.15	0.45	0.7
	1000	1	4.45	5.8	12.8	0	0	0
X3 and X2 Only	200	4	2.7	3.3	2.2	5.8	6.4	5.0
	200	1	5.8	6.4	4.3	1.3	3	5.3
	1000	4	4.7	5.3	4.7	0.4	0.7	1.4
	1000	1	3.5	5.3	6.7	0	0	0
Null Model	200	4	39.2	30.4	41.4	7.0	6.6	11
	200	1	20.4	12.8	35.6	0.1	0.3	1.6
	1000	4	17.5	15.0	21.4	0	0	0
	1000	1	0.4	1.4	4.7	0	0	0

N^{SC} Subcohort size; $\frac{N^{NC}}{N^{C}}$ non-case to case ratio

6.4.5.3 Sparse Data

Table 6.6 shows the effect of sparse data on the performance of hypothesis tests in the full cohort and case cohort. As in Table 6.3, the mean of power for all combinations of "false" parameters $\gamma 2$ and $\gamma 4$ included with specific combinations of "true" test parameters $\gamma 1$ and $\gamma 3$ in the test parameters is presented. In this table, column "false" presents the number of such false parameters included in the test parameters. For clarity and brevity, the mean Type 1 error rate over all combinations of false parameters $\gamma 2$ and $\gamma 4$ is shown. Recall that covariate $X3$ was generated with extremely sparse data as might yield a non-normal likelihood in the full cohort, with only 2 events with $X3 = 1$ in each replicate.

Type 1 error rate is similar for all tests. LR and dLR behave similarly to 6.3 where the DGM was not designed to generate sparse data. The Wald tests, by contrast, show lower power than LR and dLR when $\gamma 3$ is included in the test parameters. This is most apparent at the lower HR/SD for $X1$, as the high power in the full cohort when HR/SD for $X1 = 1.3$ obfuscates this result for the full cohort Wald test.

Table 6.6: Sparse Data: Performance of Hypothesis Tests of Parameters in M_M

(Staggered Entry, sampling fraction 0.03, non-case to case ratio 1 and IPW Classic)

	HR/SD (X1):		1.1				1.3			
	Test Parameters		Full Cohort		Case Cohort		Full Cohort		Case Cohort	
	true	false	Wald	LR	Wald	dLR	Wald	LR	Wald	dLR
Type 1 Error	0	>0	5.4	5.4	5.6	5.6	5.1	5.1	5.5	5.6
Power	$\gamma 1$ $\gamma 3$	0	94.5	100	66.1	80.0	100	100	99.9	99.9
		>0	90.8	99.9	60.5	72.1	100	100	99.8	99.9
	$\gamma 1$	0	72.7	72.8	43.6	43.8	100	100	99.9	99.9
		>0	70.3	70.3	39.2	39.3	100	100	99.7	99.7
	$\gamma 3$	0	100	100	53.9	81.5	100	100	55.9	79.9
		>0	32.6	99.8	30.1	52.0	34.2	99.8	31.1	51.0

Table 6.7 shows the effect of sparse data on AIC, BIC, and their case-cohort modifications. While pBIC, pAIC and dAIC show similar behavior as in the previous simulation study, dBIC is less likely than BIC or $dBIC$ to select models that contain $X3$, the covariate with non-normal profile likelihood, resulting in low power to select the true model.

Table 6.7: Sparse Data: Model Selection

(Staggered Entry, sampling fraction 0.03, non-case to case ratio 1 and IPW Classic)

HR/SD (X1)	Model	AIC	pAIC	dAIC	BIC	pBIC	dBIC
1.1	**True Model**	**65.2**	**34.1**	**46.8**	**63.1**	**33.2**	**9.4**
	True Model + X2 or X4	11.3	16.75	7.95	0.35	2.3	0.95
	Univariate Model X1	0	1.1	2.7	0.2	15.5	39.2
	Univariate Model X3	4.7	6.5	16.3	32.8	23.4	6.7
	X3 and X2 Only	3.3	6.6	7.9	3	5	1.6
	Null Model	0	0.6	1.8	0	11.2	25.5
1.3	**True Model**	**71**	**43.7**	**65**	**97.9**	**57.5**	**18.3**
	True Model + X2 or X4	13.25	20.85	12.6	1.05	4.5	1.85
	Univariate Model X1	0	1.9	5.1	0	28.1	65.1
	Univariate Model X3	0	0	0	0	0	0
	X3 and X2 Only	0	0	0.1	0	0.1	0.1
	Null Model	0	0	0	0	0	0

6.5 Discussion

I first discuss the results where the data generating mechanism was not designed to introduce sparse data, followed by the impact of sparse data, and an overall conclusion.

6.5.1 Hypothesis Testing

In the full cohort, the Wald and Likelihood Ratio tests generally behaved as expected. They showed Type 1 error rates close to 5%, and similar power to each-other. In the case-cohort, robust Wald tests showed similar Type 1 error rates to the full cohort, and power was diminished from the full cohort to a similar degree as seen in comparable scenarios in Chapter 3. Loss of power diminished as subcohort size and number of cases in the dataset increased. The na$\ddot{}$ieve pLR showed high Type 1 error rates $> 10\%$. The dLR modification, however, shows similar results to the Wald tests. Overall, this simulation study confirms the work of Lumley & Scott in the application of dLR to the case-cohort design.

6.5.2 Model Selection

In the case cohort, $pBIC$ and $dBIC$ perform similarly to each-other, and indeed, Lumley & Scott note in their 2015 paper that $pBIC$ and $dBIC$ are likely to select the same model in many scenarios. The application of $dAIC$ to case-cohort data in

these simulation studies showed that $dAIC$ selected models more similarly to AIC in the full cohort than did $pAIC$. The rate of selection of the true model was low for both full cohort and case cohort when β was small and/or number of cases was low. As the subcohort size and number of cases increased, loss of power from the full cohort in the case cohort diminished.

6.5.3 Impact of Sparse Data

In this study, the data-generating mechanism created extremely sparse data, and results may not be as dramatic in real world applications. However, in this simulation study, Wald tests showed substantially reduced power when the test parameters included the γ corresponding to the covariate with sparse data. Similarly, $dBIC$ showed reduced power to select the true model when data was sparse.

6.5.4 Conclusion

The application of methods from the field of complex survey sampling has allowed for valuable methods for hypothesis testing and model selection in the case-cohort design. Overall, the results indicate that dLR and Wald tests display reasonable Type 1 error rates in the case cohort, and that power increases as subcohort size and number of cases increase. Similarly, as subcohort size and number of cases increase, $dAIC$, $pBIC$, and $dBIC$ behave mores similarly to their full cohort equivalents. However, $dBIC$ and Wald tests should be regarded with caution where data is sparse or a non-normal profile pseudolikelihood is suspected.

Chapter 7

Application to InterAct

7.1 Introduction

In this chapter the use of the methods described previously is illustrated using the InterAct dataset. In Section 7.2 I briefly describe the InterAct Case-Cohort Study. In Section 7.3 I describe the InterAct Consortium et al. (2012a) paper, in which both the independent association of Physical Activity with incident type 2 diabetes and the association between Physical Activity and diabetes incidence within strata of BMI or WC were assessed separately in men and women. In this section I also discuss measures of obesity, and the role of obesity in the causal pathway of type 2 diabetes. In Section 7.4 I describe the aims of my analysis, which are to explore what the methods of this thesis can tell us about the estimation, selection and criticism of a selection of the models fitted by InterAct Consortium et al. (2012a). In Section 7.5 I describe the subset of the InterAct data analysed in this chapter. The subject-handling centres over which models were estimated are summarised in Table 7.1. In Table 7.2 I summarize the covariates included in the analysis dataset.

In Section 7.6 I fit an initial model to the analysis dataset, similar to that of InterAct Consortium et al. (2012a) but with a stratified Cox model rather than combining the effects using random effects meta-analysis. In Section 7.7 I choose functional forms for Physical Activity and the other continuous covariates in each sex, with the aim of capturing any non-linear effects. In Section 7.8 I test for violation of the assumption of proportional hazards, and include a time-varying effect where violations are detected. In Section 7.9 I fit additional models by similar methods, each including a different measure of obesity (Waist Circumference and Waist to Height Ratio). In each sex, I then use model selection methods to choose between these 3 different obesity-measure-models. Finally, in Section 7.10 I discuss the analysis and its results, and offer a final conclusion.

I find that while results are slightly attenuated in the analysis of this chapter, the results of InterAct Consortium et al. (2012a) in their investigation of the independent association of Physical Activity with incident type 2 diabetes are overall robust to their model assumptions, but that improved functional form of covariates attenuates the effect. I also find that the model with Waist-to-Height ratio is strongly preferred in both sexes. I therefore conclude that, in this analysis, a one-level increase in Physical Activity is independently associated with a relative risk reduction of 6% in Males (95% CI 1% to 10%) and 5% in Females (95% CI 0% to 9%).

7.2 The InterAct Case-Cohort Study

The InterAct study is a large prospective case-cohort study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC), designed to allow for examination of genetic and lifestyle factors on incidence of type II diabetes. The participants, methods, study design and measurements are described in detail in InterAct Consortium et al. (2011). EPIC collected standardised information on lifestyle exposures, socio-economic status, education, and occupation in 519,978 participants and 10 countries (Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden & the United Kingdom). All but Norway and Greece participated in InterAct, for a total of 455,680 participants in 26 subject-handling centres (within-country study locations). 109,625 individuals without stored blood and 5,821 individuals without information on diabetes status were excluded, resulting in 340,234 participants eligible for inclusion.

A random sample of 16,835 individuals was selected from the 346,055 participants with stored blood, stratified by centre. Of these, 548 individuals with prevalent diabetes, 129 individuals without information on diabetes status, and 4 individuals with post-censoring diabetes were excluded, resulting in a subcohort of size 16,154. The final InterAct dataset consists of 27,779 individuals, consisting of 15,376 non-cases, 778 subcohort cases, and 11,625 non-subcohort cases, drawn from a full cohort of 340234 with 3.65% cases. The overall sampling fractions were therefore 4.75% and 4.69% for all subcohort members and subcohort non-cases, respectively, with non-case to case ratio in the case cohort sample 1.24:1.

Male subjects are not present in French centres, the Naples centre, and the Utrecht centre. Male subjects consist of 11,892 total subjects, consisting of 5,727 non-cases, 384 subcohort cases and 5,781 non-subcohort cases, drawn from a full cohort of 125,233 subjects with 4.92% cases. The overall sampling fractions are therefore

4.88% and 4.81% for all subcohort members and subcohort non-cases, respectively, with non-case to case ratio in the case cohort sample 0.93:1. Female subjects are present in all centres. Female subjects consist of 15,887 total subjects, consisting of 9,649 non-cases, 394 subcohort cases and 5,844 non-subcohort cases, drawn from a full cohort of 215,001 subjects with 2.90% cases. The overall sampling fractions are therefore 4.67% and 4.62% for all subcohort members and subcohort non-cases, respectively, with non-case to case ratio in the case cohort sample 1.55:1. The higher non-case to case ratio seen in females than males is likely because subcohort selection was not stratified by sex, and indicates that females have lower risk.

7.3 InterAct Consortium et al. (2012a)

In this paper, two separate investigations are performed, each again performed separately in men and women. The authors note that both obesity and low levels of physical activity are important modifiable risk factors for type 2 diabetes, and "previous observational studies have suggested that higher levels of physical activity are associated with lower risk of diabetes independently of obesity".

1. For independent association of Physical Activity with incident type 2 diabetes, separate models were fitted, each considering one of two measures of obesity as a confounder; body mass index (BMI) or waist circumference (WC). They concluded that a one-category difference in physical activity was independently associated with a reduction in risk of type 2 diabetes; 13% (HR 0.87, 95% CI 0.80, 0.94) in men and 7% (HR 0.93, 95% CI 0.89, 0.98) in women.

2. Considering obesity as an effect modifier, the authors also investigate the association between PA and diabetes incidence within strata of BMI or WC. They concluded that increased physical activity was associated with a reduced risk of type 2 diabetes across all strata of BMI, and also in abdominally lean and obese men and women.

All investigations also included the following covariates as confounders; Alcohol Consumption, Energy/Calorie Intake, Smoking Status, and School Level.

In this chapter, analysis (1) above is the target analysis to which I apply the methods described in previous chapters. As obesity is itself affected by physical activity, the inclusion of a measure of obesity as a confounder means that, in this analysis, the authors were not estimating the total effect of PA, only the direct effect i.e. the part that is not mediated by BMI/WC. It is likely this analysis underestimates the total effect of physical activity on type 2 diabetes. However, given a bi-direction

association between physical activity and obesity, there is somewhat of a dilemma, as not adjusting induces confounding while adjusting ignores the mediated effect. This can be considered as an example of time-varying confounding as discussed in Daniel et al. (2013). That said, the authors appear to have taken their associations to have public health communication implications for the promotion of physical activity and, given the effect of including an obesity measure as a confounder is to underestimate the total effect of physical activity, this approach seems reasonable.

7.4 Investigation in this Chapter

In this chapter I investigate the independent association of Physical Activity with incident type 2 diabetes, fitting similar models to those used for this investigation in InterAct Consortium et al. (2012a). I use and demonstrate the case-cohort methods described previously in this thesis, and compare the results at each step with the results from this investigation in InterAct Consortium et al. (2012a).

I first fit an initial model in each sex, similar to that of InterAct Consortium et al. (2012a), with the same covariates included in the model and all continuous covariates in linear form. Keele (2010) notes that unmodeled nonlinearity can present as a violation of the proportional hazards assumption, and recommends assessment of and correction for non-linearity prior to tests for non-proportional hazards. Hence, functional form of continuous covariates is considered first. I select functional forms for the continuous covariates using the methods described in Chapters 4 and 6. Next, I assess whether the assumption of proportional hazards has been violated using the methods in Chapter 5, and include a time-varying covariate where violation is found. Finally, I select between three models using the methods of Chapter 6, each model differing by which measure of obesity is included; body mass index (BMI), waist circumference (WC), or waist to height ratio (WHtR).

(InterAct Consortium et al., 2012b) found WC and BMI to be independently associated with type 2 diabetes in the InterAct dataset (note that this is a separate investigation to the paper discussed in this Chapter, InterAct Consortium et al. (2012a)). WHtR has been increasing in popularity as a universal non-sex-specific measure (Mirzaei and Khajeh (2018), Son et al. (2016)). An alternative approach would be to include all measures as potential confounders. Measures of obesity are highly correlated, and Groenwold et al. (2016) note that where omitted confounders are correlated with included confounders, the bias caused by omitted confounders is mitigated. Additionally, reports from the literature indicate that interactions between measures of obesity may be informative (eg interactions of BMI and WC in

InterAct Consortium et al. (2012b)), however such interactions are not considered.

7.5 Analysis Dataset

In this section, I describe the subset of the InterAct dataset analysed in this chapter. As in InterAct Consortium et al. (2012a), observations missing data for Physical Activity are excluded. In Males, 96 subcohort non-cases, 6 subcohort cases, and 91 subcohort non-cases are excluded. In Females, 114 subcohort non-cases, 4 subcohort cases, and 74 subcohort non-cases are excluded. Exclusion rates vary between subject-handling centres, even within countries. A number of subject-handling centres show no exclusions in the case-cohort sample and Bilthoven, Netherlands, displays the highest rate of exclusions at $\sim$15% of non-cases.

Waist measurements for the Umea centre in Sweden are not available, hence this centre is excluded from all analyses. With the exception of Umea, the rates of missing data on Obesity Measures are low. Excluding Umea, the total number of observations with missing data for BMI, WC, WHR and WHtR, respectively, is 99, 71, 80, and 89 for females; and 63, 78, 89, and 93 for males. The final analysis dataset consists of 5,155 non-cases, 360 subcohort cases and 5,230 non-subcohort cases, drawn from a full cohort of 104,205 in males; and 9036 non-cases, 373 subcohort cases, and 5,408 non-subcohort cases, drawn from a full cohort of 192,089 in females.

Table 7.3 shows full cohort and subcohort size, subcohort non-case sampling fractions, and case-cohort non-case to case ratios for males and females in each centre, post exclusions. Due to the study design, sampling fractions and non-case to case ratios vary between subject-handling centres. However, reflecting the study as a whole, sampling fractions within centres are generally slightly larger for males than females, and non-case to case ratios are generally larger for females than males.

Physical Activity is categorical with 4 levels; inactive, moderately inactive, moderately active, and active. Smoking Status is categorical with 3 levels; never smoked, past smoker, and current smoker. School Level is categorical with 5 levels; none, primary school, technical school, secondary school, and post-secondary school. School level None is not present in Cambridge, Oxford, Netherlands, or Denmark, and amounts to less than 20 observations for all other locations except Spain. Hence, School level None and Primary are combined for this analysis. Frequencies for categorical covariates are shown in Table 7.1. BMI, WHR, WC, Alcohol Consumption and Calorie Intake are recorded as continuous covariates. WHtR was calculated from height and waist measurements. Summary statistics for continuous covariates

are shown in Table 7.2. In tables, Alcohol is presented as g/day$\times 10^{-1}$; Calories as kcal/day$\times 10^{-3}$; WHtR as WHtR$\times 10^{2}$; and WHR as WHR$\times 10^{2}$

Table 7.1: InterAct Study: Frequencies of Categorical Covariates

		Male			Female		
		Non-Case	SC Case	Non-SC Case	Non-Case	SC Case	Non-SC Case
PA	Inactive	905	71	1285	2400	142	1887
	Mod. Inactive	1623	124	1702	3223	123	1858
	Mod. Active	1288	92	1144	1867	59	964
	Active	1339	73	1099	1546	49	699
School	None/Primary	2039	188	2568	3708	224	2945
	Technical	1155	80	1206	2091	89	1208
	Secondary	599	44	500	1447	20	611
	Further	1307	46	894	1658	35	525
Smoke	Never	1549	78	1192	5015	220	3071
	Former	1918	153	2193	1968	68	1164
	Current	1672	128	1829	2007	85	1140

Table 7.2: InterAct Study: Summary Statistics for Continuous Covariates

		Male			Female		
		Non-Case	SC Case	Non-SC Case	Non-Case	SC Case	Non-SC Case
Calories	Mean	2.50	2.48	2.47	1.95	1.87	1.94
	SD	0.64	0.66	0.67	0.52	0.55	0.55
	Median	2.44	2.40	2.39	1.89	1.77	1.86
Alcohol	Mean	2.34	2.33	2.42	0.83	0.65	0.67
	SD	2.40	2.50	2.67	1.18	1.06	1.09
	Median	1.61	1.51	1.54	0.37	0.16	0.18
BMI	Mean	26.54	29.71	29.38	25.54	30.61	30.10
	SD	3.42	3.56	4.03	4.32	5.41	5.32
	Median	26.23	29.45	29.01	24.84	30.12	29.51
WHtR	Mean	54.46	60.09	59.30	50.12	58.98	57.98
	SD	6.03	5.78	6.26	7.25	8.27	8.04
	Median	54.12	59.86	58.83	49.07	59.45	57.59
WC	Mean	94.53	103.19	102.54	80.67	94.15	92.66
	SD	9.76	9.39	10.47	10.78	12.38	12.23
	Median	94.00	102.70	102.00	79.00	94.00	92.00

Note: Alcohol g/day$\times 10^{-1}$; Calories kcal/day$\times 10^{-3}$; WHtR$\times 10^{2}$; WHR$\times 10^{2}$

Table 7.3: InterAct Study: Full Cohort (FC) and Subcohort (SC) Size, Sampling Fractions, and Non-Case to Case Ratio Post Exclusions

Country	Centre	Male						Female					
		Non-Case N		Case N		α^{NC}	$\frac{N^{NC}}{N^{C}}$	Non-Case N		Case N		α^{NC}	$\frac{N^{NC}}{N^{C}}$
		FC	SC	SC	Non-SC			FC	SC	SC	Non-SC		
France	Ile-de-France	0	0	0	0			4961	144	3	65	2.90	2.12
	North-West							3724	110	1	52	2.95	2.08
	North-East							3703	112	3	71	3.02	1.51
	Rhone-Alpes							3172	89	1	42	2.81	2.07
	Provence							2567	68	1	31	2.65	2.13
	South-West							1995	56	0	18	2.81	3.11
	Total							20122	579	9	279	2.88	2.01
Italy	Florence	3210	122	5	129	3.80	0.91	9421	408	8	255	4.33	1.55
	Varese	2347	72	3	51	3.07	1.33	9017	282	6	186	3.13	1.47
	Ragusa	2522	144	8	139	5.71	0.98	3001	172	6	104	5.73	1.56
	Turin	5319	289	12	190	5.43	1.43	4078	223	3	85	5.47	2.53
	Naples							4726	213	9	184	4.51	1.10
	Total	13398	627	28	509	4.68	1.17	30243	1298	32	814	4.29	1.53
Spain	Asturias	2613	279	28	256	10.68	0.98	4984	451	33	240	9.05	1.65
	Granada	1312	102	7	93	7.77	1.02	4583	395	25	193	8.62	1.81
	Murcia	1996	224	18	208	11.22	0.99	4739	491	28	295	10.36	1.52
	Navarra	3196	328	40	301	10.26	0.96	3636	377	31	209	10.37	1.57
	San Sebastian	3341	306	35	341	9.16	0.81	3791	372	18	165	9.81	2.03
	Total	12458	1239	128	1199	9.95	0.93	21733	2086	135	1102	9.60	1.69
UK	Cambridge	9002	376	13	419	4.18	0.87	11021	507	13	298	4.60	1.63
	Oxford	3318	92	0	101	2.77	0.91	9563	246	1	136	2.57	1.80
	Total	12320	468	13	520	3.80	0.88	20584	753	14	434	3.66	1.68
Netherlands	Bilthoven	7175	233	6	136	3.25	1.64	8492	243	7	110	2.86	2.08
	Utrecht							15722	895	25	481	5.69	1.77
	Total	7175	233	6	136	3.25	1.64	24214	1138	32	591	4.70	1.83
Germany	Heidelberg	9608	373	16	472	3.88	0.76	11465	465	16	276	4.06	1.59
	Potsdam	9093	456	12	454	5.01	0.98	14343	700	16	322	4.88	2.07
	Total	18701	829	28	926	4.43	0.87	25808	1165	32	598	4.51	1.85
Sweden	Malmo	9644	707	69	832	7.33	0.78	15709	1,090	60	791	6.94	1.28
	Umea	10973	476	18	460	4.34	1.00	11974	499	17	362	4.17	1.32
	Total	20617	1183	87	1292	5.74	0.86	27683	1589	77	1153	5.74	1.29
Denmark	Aarhus	7856	327	22	356	4.16	0.87	8288	297	18	244	3.58	1.13
	Copenhagen	17063	725	66	752	4.25	0.89	19607	630	41	555	3.21	1.06
	Total	24919	1052	88	1108	4.22	0.88	27895	927	59	799	3.32	1.08
Total (excl. Umea)		98615	5155	360	5230	5.23	0.92	186308	9036	373	5408	4.85	1.56

$\frac{N^{NC}}{N^{C}}$ non-case to case ratio; α^{NC} non-case sampling fraction (%)

7.6 Initial Stratified Cox Model

I now describe in detail the procedures for analysis of models including BMI as the measure of obesity, followed by model selection between models with alternate measures of obesity (the results for each step for these models are included in the appendix. Table 7.5 summarizes estimated HR for Physical Activity at each step of the analysis, Compared with those from InterAct Consortium et al. (2012a).

Table 7.4: Hazard Ratios for Physical Activity - Comparison of Analysis Steps

Table 7.5

Model	Men		Women	
	HR	95% CI	HR	95% CI
Interact 2012 (BMI)	0.87	0.80, 0.94	0.93	0.89, 0.98
Startified Cox Model (BMI)	0.88	0.84, 0.93	0.93	0.89, 0.97
Improved Functional Form (BMI)	0.90	0.86, 0.94	0.94	0.90, 0.98
Included Non-Proportional Hazards (BMI)	0.90	0.86, 0.94	0.94	0.89, 0.98
WHtR Model	0.94	0.90, 0.99	0.95	0.91, 1.00

Outcome: Type 2 Diabetes Incidence

WHtR Model following Improved Functional Form and Inclusion of Non-Proportional Hazards

All Models Include Physical Activity, Calories, Smoking, School, Alcohol, BMI

In InterAct Consortium et al. (2012a) seperate models were fitted in each location (centre) and the effects combined using random effects meta-analysis. In all my analyses, stratified Cox models are fitted using stratified IPW weighting. In the context of the Cox model, stratified Cox models refer to Cox models where coefficients are equal across strata but baseline hazards are allowed to vary across strata. In this thesis, this procedure is referred to as stratified modelling. In all analyses, stratified modelling is carried out with 10 potential strata; countries: Denmark, France, Italy, Spain, the Netherlands, Sweden; individual study locations: Germany - Potsdam, Heidelberg; UK - Oxford, Cambridge. As males are absent from French centres, males are modelled with 9 strata. Subject-handling centres in Germany and the UK are included as individual strata due to minor differences in measurements of Obesity Measures and Physical Activity between centres in these countries.

In all analyses, age is used as time-scale, and Huber sandwich estimates are used for calculation of coefficient standard errors, confidence intervals, and p-values. IPW weighting is chosen rather than Prentice due to easier implementation of the methods in STATA with this weighting method. All covariates were centered at the IPW-weighted mean of the case-cohort sample for each sex. The baseline hazard function can hence be interpreted as the hazard function for an individual with "av-

erage" values for the covariates. Post-Secondary/Further was used as the reference category for School, and None was used as the reference category for Smoking. HRs for Physical Activity following this step are: 0.88 (95% CI 0.84 to 0.93) in men and 0.92 (95% CI 0.89 to 0.97) in women. This is similar to the InterAct Consortium et al. (2012a), though slightly attenuated in men (see Table 7.5).

7.7 Functional Form of Covariates

In InterAct Consortium et al. (2012a) and in this analysis, Physical Activity (PA) was modelled as a linear effect. As PA contains only 4 levels, smooths of martingale residuals are uninformative and not presented. Fig 7.1 shows smooths of martingale residuals against continuous covariates for; models where all continuous covariates are in linear form, and in final functional form. Smooths were fitted at 1000 quantiles of the IPW-weighted covariate values, using the default Epanechnikov kernel, and degree 0. Bandwidth was set as 1/50th the range of the covariate. Whether a non-linear relationship might be appropriate is first assessed by inspection of smooths of martingale residuals in combination with reports from the literature. Where a non-linear relationship is indicated, it is modelled by restricted cubic splines, which produce a continuous smooth function, linear in the tails and a piecewise cubic polynomial between adjacent knots (Croxford, 2016). Choice of knots for each covariate was guided by the appearance of smooths of martingale residuals, commonly used benchmarks, and reports from the literature. In Section **??** results for statistical assessment of non-linearity for these covariates are presented.

7.7.1 Calories

Smooths of martingale residuals against Calories appeared broadly linear in the bulk of the data, with some fluctuations but without obvious trend. The most obvious deviations were in the tails. Smooths of martingale residuals for evaluation of functional form are known to be unreliable in the tails (Ganguli et al., 2015), particularly where sample size is small, due to additional variation in cumulative baseline hazard from reduction in the effective sample size by prior failure and censoring. Note that in the location-stratified Cox model, cumulative baseline hazard, and hence martingale residuals are calculated within each stratum. The risk sets over which martingale residuals are calculated are therefore smaller than would be expected from the overall size of the subcohort. Hence individual martingale residuals may be less reliable than might be assumed. Weighting of smooths may exacerbate this effect. I did not find indications from literature that Calories has a non-linear effect on incidence of type 2 diabetes. Calories was hence modelled in linear form.

7.7.2 Alcohol Consumption

There is some discordance in the literature for influence of Alcohol consumption on risk of Type 2 diabetes. The literature variously reports reduced risk compared to light and heavy drinkers in males at 6-48g/day and females at 6-24g/day (Koppes et al., 2005), reduction of risk compared to non-drinkers and heavy drinkers in males and females at <24g/day, (Li et al., 2016), reduced risk compared to non-drinkers at alcohol consumption <71 g/day in females, and a small increase in risk as compared to non-drinkers with any alcohol consumption in males (Knott et al., 2015). Restricted cubic splines were fitted using 4 knots at prespecified locations according to the percentiles of the distribution of Alcohol, the 5th, 25th, 75th, and 95th percentiles.

7.7.3 BMI

Smooths of martingale residuals against BMI displayed J-shaped curves. A non-linear relationship of certain measures of obesity with diabetes and metabolic syndrome has been previously described (e.g. Su et al. (2016), Yu et al. (2018)). BMI has two commonly used boundaries for increased health risk from the literature, with BMI 25 and 30 corresponding to Overweight and Obese in WHO Guidelines (WHO, 2008). Restricted cubic splines were fitted using 4 knots at prespecified locations according to the percentiles of the distribution of BMI, the 5th, 25th, 75th, and 95th percentiles.

7.7.4 Statistical Assessment of Non-Linearity

I next use variable selection methods to assess whether use of restricted cubic splines for the continuous confounders improves model fit. While in practice a particular variable selection method or combination of methods might be chosen, here results are presented for Wald, dLR, $dAIC$, $dBIC$ and $pBIC$ to demonstrate similarities and differences between the available methods. Further, these methods could be applied in a more formal selection procedure such as backwards stepwise selection.

Table 7.6 shows the results for the significance of the use of restricted cubic splines by Wald Tests and dLR. In each test, the full model includes Physical Activity as a linear effect, Calories in linear form, Smoking Status and School level as categorical covariates, restricted cubic splines for Alcohol and restricted cubic splines for BMI. The test parameters are the addition variables created by the restricted cubic spline procedure. For example, the result of p=0.093 for Wald in row 2 indicates that there is no evidence to reject the null hypothesis that the effect of including

restricted cubic splines for Alcohol is equal to 0. Table 7.6 also shows the difference in $dAIC$, $dBIC$ and $pBIC$ between a model where the covariate of interest is modelled in linear form, and a model where the covariate of interest is modelled with restricted cubic splines. All models also include Calories in linear form, Smoking Status, School Level, Physical Activity as a linear effect, and the non-linear form of whichever of Alcohol or the relevant Obesity Measure is not the covariate of interest. For example, the result of -1.95 for $dBIC$ in row 2 is the difference in $dAIC$ between a model in males with Physical Activity as a linear effect, Calories in linear form, Smoking Status, School Level, restricted cubic splines for BMI and a linear effect for Alcohol; and a model that is the same in all respects except that it includes restricted cubic splines for Alcohol. As the result is negative, we hence prefer the model with restricted cubic splines for Alcohol. Note that dBIC for the maximal model will be equal to 0, and hence where dBIC is greater than 0, the restricted cubic splines are preferred.

Table 7.6: InterAct Study: Assessment of Model Fit with Restricted Cubic Splines

		dAIC Difference from Linear Form	pBIC Difference from Linear Form	dBIC	Wald	dLR
Male	BMI	-227.60	-196.01	154.28	<0.001	<0.001
	Alcohol	-1.95	3.52	-1.05	0.093	0.094
Female	BMI	-600.61	-565.27	488.50	<0.001	<0.001
	Alcohol	-47.82	-39.60	40.83	<0.001	<0.001

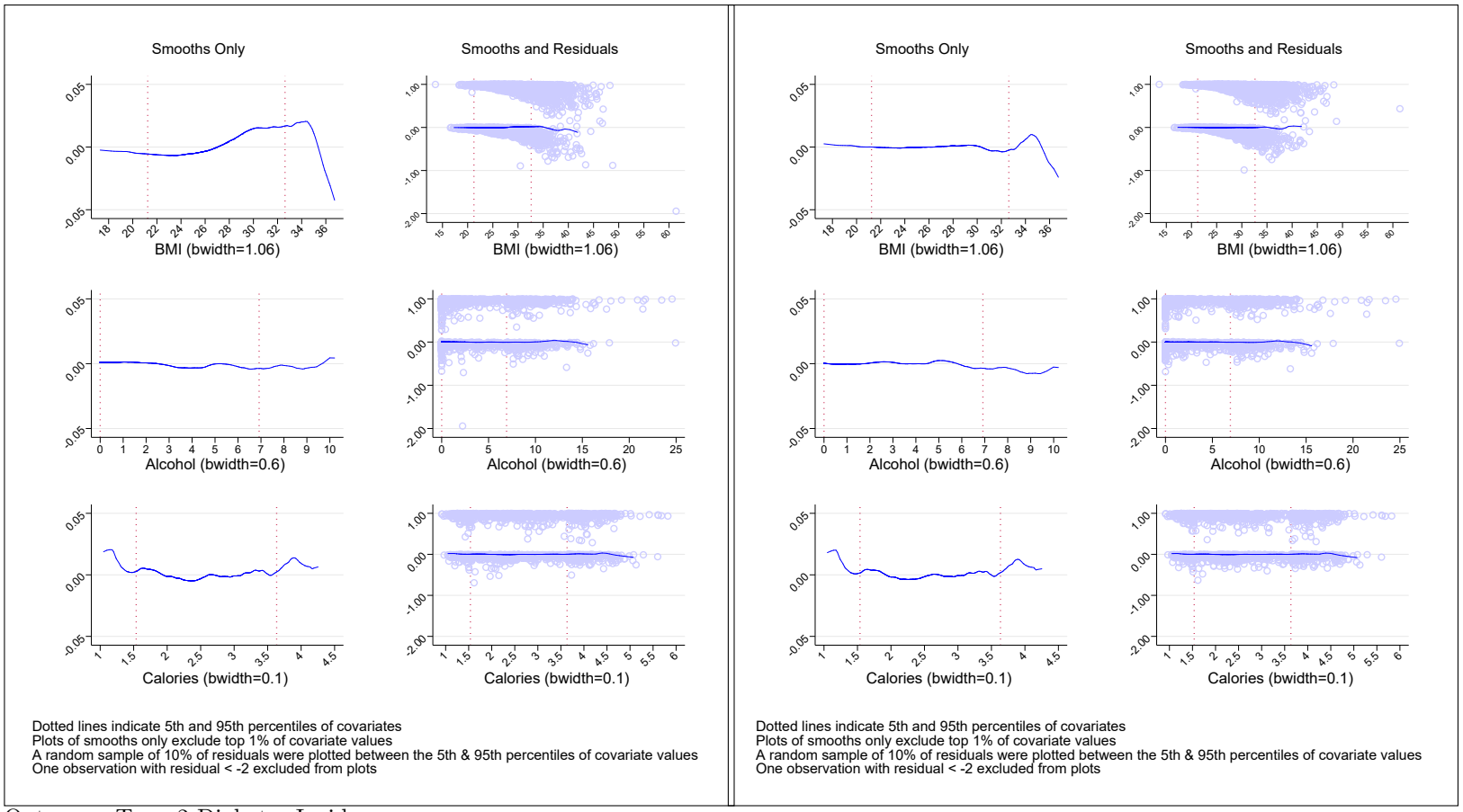
Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

In both sexes, restricted cubic splines for BMI improves model fit, as indicated by both significance testing and the information criteria. In Females, all assessments indicate that restricted cubic splines for Alcohol improve model fit. In Males, $dAIC$ indicates that the use of restricted cubic splines for Alcohol improves model fit, though with a small magnitude. As unmodelled non-linearity can affect performance of tests for non-proportional hazards, a less parsimonious approach was taken and the linear splines for Alcohol were retained in the model for Males. In summary, in both men and women, models are taken forward with restricted cubic splines for BMI, restricted cubic splines for Alcohol, and Calories as a linear effect.

Hazard ratios for Physical Activity following this step are as follows: 0.90 (95% CI 0.86 to 0.94) in men and 0.94 (95% CI 0.90 to 0.98) in women, a further attenuation of effect from the previous step.

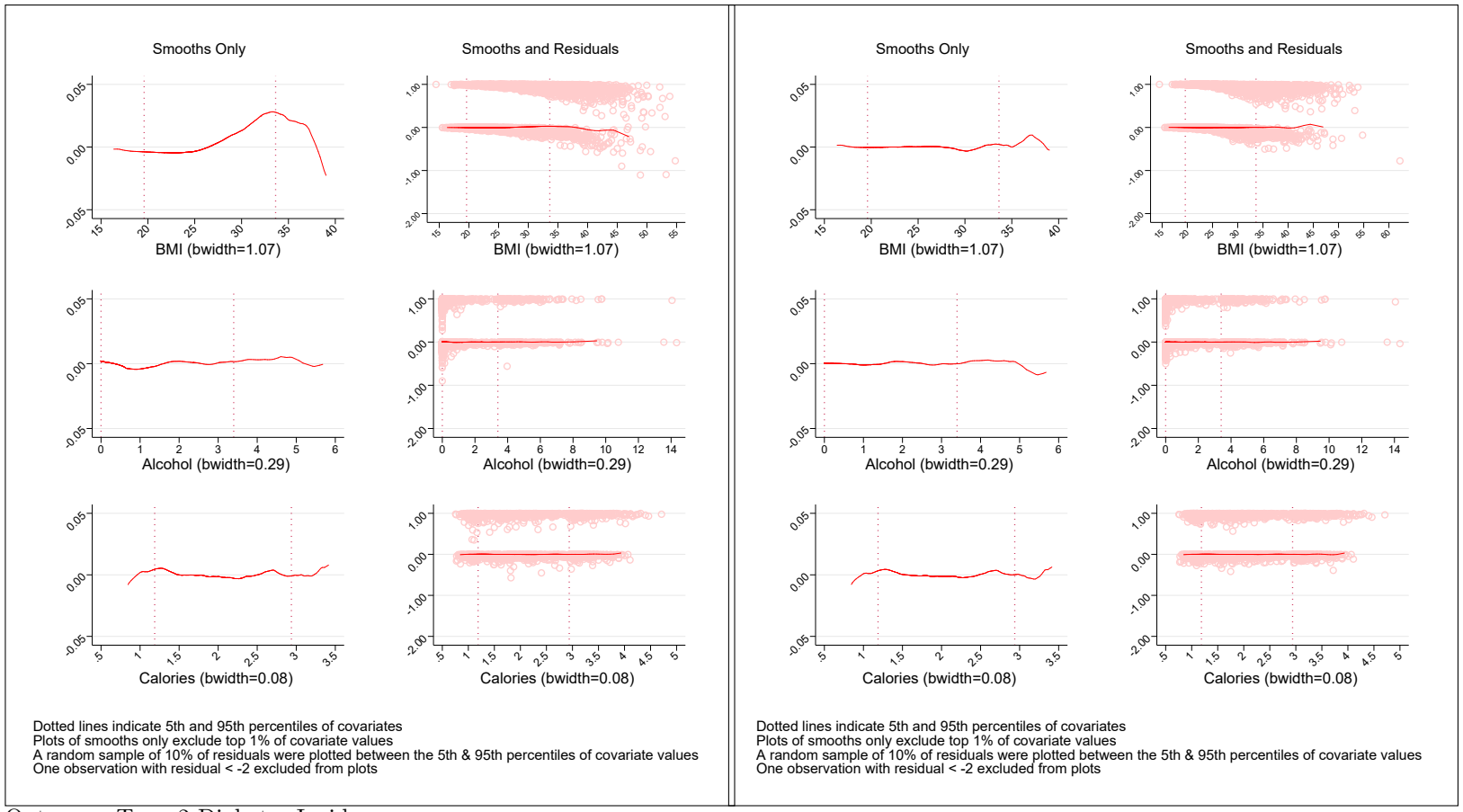
Figure 7.1: InterAct Study: Smooths of Martingale Residuals in Men



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

Figure 7.2: InterAct Study: Smooths of Martingale Residuals in Women



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

7.8 Detection of Non-Proportional Hazards

Using the functional forms chosen in Section 7.7, the models are re-fitted with all covariates simultaneously allowed to interact with rank of time. For each covariate, Wald tests are carried out for the null hypothesis that for all terms representing that covariate, interactions with rank of time are simultaneously equal to 0. For example, for School Level, the test parameters are the interactions of all levels of School with rank of time.

As described in Chapter 6, dLR can be more reliable than Wald tests where data is sparse or a non-normal profile pseudolikelihood is suspected. An advantage of the Wald test is that multiple test parameters can be assessed after fitting a single model. To assess the same test parameters with dLR requires fitting of the maximal model twice (to retrieve robust and naieve variance estimates), and fitting each relevant submodel, again twice. Inclusion of time-varying covariates is computationally intensive. Given the computational load of calculating dLR for two sexes, and each of those with 6 covariates, dLR is not assessed.

Table 7.7 shows the resulting p-values for each covariate. Covariates with p-values below a threshold of 0.05 are considered to violate the assumption of proportional hazards. In both men and women, BMI displays evidence of non-proportional hazards. The remaining covariates do not display evidence of non-proportional hazards. Going forward, I adjust both models to account for these effects by including interactions with rank of time for all terms representing BMI in the model.

Table 7.7: InterAct Study:p-values for Wald Tests for Interaction of Covariates with Rank Time

	PA	BMI	Alcohol	Calories	School	Smoke
Men	0.462	<0.001	0.671	0.445	0.819	0.891
Women	0.360	<0.001	0.437	0.478	0.411	0.154

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

Hazard ratios for Physical Activity following this step are as follows: 0.90 (95% CI 0.86 to 0.94) in men and 0.94 (95% CI 0.89 to 0.98) in women, essentially unchanged from the previous step.

7.9 Model Selection

The procedure above was repeated for models that substituted an alternative obesity measure (WHtR and WC) for BMI. Final functional forms were the same as those for BMI. Evidence of violation of the assumption of proportional hazards was found for the covariate measuring obesity. Martingale residual smooths, Tables of assessment of model fit, and results for Wald tests for Interaction of covariates with rank time for these models can be seen in the Appendix. The resulting final models in each sex and for each Obesity Measure are now summarized: All models include the categorical covariates School Level and Smoking Status. All models include Physical Activity and Calories as linear effects. All models include Alcohol and the relevant Obesity Measure as restricted cubic splines with 4 knots as described in Section 7.7. All models include the restricted cubic splines for the relevant Obesity Measure interacting with rank of time.

Table 7.8 shows Hazard Ratios, standard errors, p-values and 95% confidence intervals in each model for the effect of a one level difference in Physical Activity on the incidence of type 2 diabetes, together with the difference in $dAIC$ and $pBIC$ from the preferred model in each sex. Ranking of models was the same for both $dAIC$ and $pBIC$. In both sexes, WHtR was the preferred model. I therefore conclude that, in this analysis, a one-level increase in Physical Activity is independently associated with a relative risk reduction of 6% in Males (95% CI 1% to 10%) and 5% in Females (95% CI 0% to 9%). In Females, this reduction in relative risk is not statistically significant at a threshold of p = 0.05.

Table 7.8: InterAct Study: Estimation Results for the Independent Association of Physical Activity and Incidence of Type 2 Diabetes

	Males			Females		
Model:	WHtR	BMI	WC	WHtR	BMI	WC
HR	0.94 (6%)	0.90 (10%)	0.95 (5%)	0.95 (5%)	0.94 (6%)	0.95 (5%)
95% CI	0.90, 0.99 (1%, 10%)	0.86, 0.94 (6%, 14%)	0.91, 0.99 (1%, 9%)	0.91, 1.00 (0%, 9%)	0.89, 0.98 (2%, 11%)	0.91, 1.00 (0%, 9%)
s.e.	0.022	0.020	0.022	0.023	0.022	0.023
p	0.006	<0.001	0.022	0.060	0.007	0.037
pBIC diff	0	295	284	0	649	166
dAIC diff	0	295	287	0	639	159

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, Obesity Measure

Values in brackets percentage reduction in relative risk

7.10 Discussion

7.10.1 Results for Different Obesity-Measure Models

In this investigation, the model using waist to height ratio as the measure of obesity is strongly preferred in both males and females. Both $dAIC$ and $pBIC$ give the same rankings, with similar intervals between criteria for the various models. In males, the BMI model produces the largest point estimates of the hazard ratio for Physical Activity, while WHtR and WC are similar. In females, the point estimates are broadly similar in all models. Standard errors for Physical Activity are also similar for each obesity-measure-model and each sex.

It is also interesting to compare the results of the models with BMI and WC to the results of InterAct Consortium et al. (2012a) upon which this study was based. For both Obesity Measures, and in both sexes, results are somewhat attenuated in this chapter as compared to InterAct Consortium et al. (2012a). In the paper, the same confounding variables are used as in this analysis, with the exception that, in this analysis, School Level None and Primary were combined. Further, in this chapter, the Umea centre in Sweden was excluded as waist measurements for this centre are not available. Also, in the paper, hazard ratios are calculated separately in each location with Prentice weighted Cox modelling, and then combined using random effects meta-analysis. By contrast, in this analysis hazard ratios are calculated using stratified IPW Classic weighting and location-stratified Cox modelling. Both the initial models fit and the change of functional form in this chapter show a small degree of attenuation in Hazard Ratio as compared to InterAct Consortium et al. (2012a), with adjustment for non-proportional hazards not an influence. It appears that overall, the results of InterAct Consortium et al. (2012a) in their investigation of the independent association of Physical Activity with incident type 2 diabetes are robust to their model assumptions, but that improved functional form attenuates the effect of Physical Activity.

7.10.2 Functional Forms of Covariates

While the appearance of the smooths of the martingale residuals for obesity measures were improved following the use of cubic splines, there was still some visual divergence from linearity at the higher values of the covariate. Changes to number of knots and knot placement were considered, but did not result in an improvements to the visual smooth. It should be noted that alternative methods for modelling of non-linear relationships, such as linear splines, may result in different results from those presented here. Restricted cubic splines have the disadvantage that coeffi-

cients are not easily interprable. However, all models included interactions of the Obesity Measure with rank of time, which would also impact on the interprability of the coefficients.

7.10.3 Non-Proportional Hazards

Keele (2010) notes that, in addition to unmodelled non-linearity, unmodelled interactions and missing covariates can also present as a violation of the proportional hazards assumption. Family history of diabetes is known to be a strong risk factor (InterAct Consortium et al., 2013), but is not included in this analysis as data for this covariate is not available for Italy, Spain, Oxford, and Heidelberg. These analyses did not consider interactions, and it is possible that inclusion of interactions might result in alterations to the tests for non-proportional hazards. For example, one might expect that interactions of Calories with Obesity and/or Physical Activity at baseline might serve as a measure of weight loss or weight gain over the course of the study. Additionally, only a single measure of Obesity is considered as a confounder in each model. InterAct Consortium et al. (2012a) find that inclusion of both WC and BMI and their interaction is informative in modeling risk of diabetes. Finally, it is of course possible that the methods and procedures used here to model non-linearity do not accurately capture the non-linear relationships that are present.

The lack of a method to apply tests for detection of non-proportional hazards that are based on Schoenfeld residuals to case-cohort data was keenly felt in the analysis of this data. Inclusion of a single interaction of a covariate with time took approximately 200 times longer in computational time than the same Cox model without the interaction. Inclusion of all covariates interacting with time took approximately 600 times longer than the same Cox model without the interactions. Calculations of Schoenfeld residuals and their associated tests is far less computationally intensive. Further, visual inspection of Schoenfeld residuals against various functions of time can assist in guidance of choice of function of time to use for formal statistical tests. In this analysis, rank of time is used for the interaction based on the recommendation of Park and Hendry (2015).

7.10.4 Stratified Modelling

Little adjustment was required in programming to apply the methods described in previous chapters to stratified modelling with IPW weighting. Prentice weighting would have required manual calculation of martingale residuals within each strata. It should be noted, however, that application of the methods used in these analysis to stratified modelling has not been studied by simulation in this thesis.

7.10.5 Conclusion

The epidemiological goal of this chapter was to estimate the independent effect of Physical Activity on the risk of incident type 2 diabetes in the InterAct dataset, adjusting for Smoking Status, Education Level, Calorie Intake, Alcohol Consumption and Obesity. As noted in InterAct Consortium et al. (2012a), Physical Activity and Obesity Measures were recorded at baseline, and this analysis does not account for any changes in Physical Activity and/or obesity between baseline and follow-up. Adjustment for multiple measures of obesity or interactions between them was not considered in this analysis and would be a valuable avenue for further research. It is possible that further improvements to functional form of the covariates or inclusion of interactions could further attenuate the independent effect of Physical Activity on type 2 diabetes incidence, and this is also a valuable area for future research. Use of model selection methods strongly preferred waist to height ratio over other Obesity Measures in both men and women. Overall, the results indicate that, independent of obesity, a one-level increase in Physical Activity is independently associated with a reduction of risk of incident type 2 diabetes by 6% in men and 5% in women.

The methodological goals of this chapter were to apply the methods described in previous chapters to a real-world dataset. The applications of the methods to IPW-weighted stratified Cox models using STATA were not difficult in terms of programming. Martingale residuals were informative in assessment of functional form. Case-cohort variable and model selection methods allowed for statistical assessment of non-linearity, and choice between candidate models. Detection of violations of the proportional hazards assumption using interactions of covariates with time was computationally intensive in this large dataset, highlighting the value of further investigations of the application to the case-cohort design of methods incorporating Schoenfeld residuals.

Chapter 8

Discussion

8.1 Introduction

In this final chapter I first summarize the key points from each preceding chapter. I next place my contributions in context by discussing the key themes considered in this thesis. Next, I consider the limitations of this thesis and discuss possible areas for future work. Finally I give a conclusion.

8.2 Dissertation Summary

In Chapter 1, I first outlined the motivation of this thesis; to move beyond parameter estimation and investigate post-estimation methods that would allow for greater exploitation of the Cox proportional hazards model in the case-cohort design. I then outlined the Cox proportional hazards model, the case-cohort design, and the data structures of survival analysis relevant to this thesis. Next, I detailed the literature regarding methods in the case-cohort design, and characteristics of the case-cohort design in practice.

In Chapter 2, I described the general data-generating mechanism used for simulation studies in this thesis. This data-generating mechanism was designed to reflect the real-world applications of the case-cohort design as described in Chapter 1, while allowing for comparison of results across different combinations of factors such as non-case to case ratio and subcohort size.

In Chapter 3, I considered methods of parameter estimation under the Cox proportional hazards model in the case-cohort design. I performed a simulation study and concluded that IPW and Prentice weighting methods achieve comparable performance for estimation of β in many scenarios, with IPW displaying improved

efficiency at higher sampling fractions, smaller non-case to case ratios and where covariates are more variable. I also compared estimation of cumulative baseline hazard by Prentice and IPW methods, which has not previously been investigated, and concluded that both weighting methods are appropriate, but noted that presence of risk sets composed only of cases can introduce profound bias, particularly for IPW.

In Chapter 4, I investigated use of weighted smooths of martingale residuals against covariate values in the choice of appropriate functional form for continuous covariates. In a simulation study, statistical assessment of non-linearity of weighted linear splines was used as a proxy for subjective visual assessment of weighted smooths. Weighted smooths of martingale residuals gave depressed Type 1 error rate in the case-cohort, and there was a large loss of power from full cohort methods. It is unclear if this loss of power is due to known issues in complex survey sampling with low power of Wald tests in weighted linear regressions, or is a true reflection of a lesser usefulness of martingale residuals in the case cohort than the full cohort. I concluded that while weighted smooths of martingale residuals against covariate values may be useful in the case-cohort design, one should expect a significant loss of power from the full cohort, particularly at smaller sampling fractions and where the number of cases in the case-cohort sample is small.

In Chapter 5, I investigated methods for detection of violation of the assumption of proportional hazards. In a simulation study, I found that in the case-cohort design, methods that rely on inclusion of interactions of covariates with time give inappropriately high type 1 error rates when number of cases is small, especially when risk set sizes are also small and correlation with a covariate displaying non-proportional hazards is present. However, with larger numbers of cases ($\sim$250+) Type 1 error rate becomes more reasonable, especially where risk set sizes are larger and less variable, and power approaches that of the full cohort. I further found that methods incorporating Schoenfeld residuals give inappropriately high type 1 error in the case cohort design when risk set sizes are smaller and more variable, as may be seen with analysis under staggered entry. Further, even in fixed entry, these methods give inappropriately high Type 1 error rate with Prentice weighting when number of controls per case is low or covariates are more variable.

In Chapter 6 I investigated methods of variable and model selection, focusing on Wald tests, Likelihood Ratio tests, AIC, and BIC. In a simulation study, I investigated Wald tests and the dLR, the modified Likelihood Ratio test of Lumley and Scott (2013), and concluded that while Wald tests are generally appropriate in

the case-cohort design, dLR is a valuable alternative where sparse data is present. I showed that the $dAIC$ and $dBIC$ modifications of Lumley and Scott (2015), and the $pBIC$ modification of Xu et al. (2013), proposed for complex survey sampling, can equally be applied to the case-cohort design, but that $dBIC$ can be inappropriate if sparse data is present.

In Chapter 7 I applied all of these methods to a real-world dataset, the InterAct Case Cohort Study, and investigated the independent effect of physical activity on the risk of incident type 2 diabetes. I first described the study and the subset of data used in the analysis. I then used smooths of martingale residuals to guide initial choice of functional form for continuous covariates in each model, followed by variable and model selection methods to assess whether such functional forms improved model fit over a linear form. I then used inclusion of interactions of covariates with time to assess violations of the proportional hazards assumption in each model, and adjusted for such violations where found to finalise each model. Finally, in each sex, I used $dAIC$ and $pBIC$ to select from the three final candidate models for each sex, each model including a different measure of obesity. I found that the models including waist-to-height ratio as the measure of obesity were preferred over those including body mass index or waist circumference.

8.3 Dissertation in Context

In this thesis I have considered the validity of estimation and post-estimation procedures in the case cohort design under both staggered entry and fixed entry, and with IPW and Prentice weighting. I have used simulation studies to assess the performance of such methods in a number of scenarios, and compared these results to those of the full cohort. Overall, this work takes place in a context where there has been little investigation of post-estimation methods in the case-cohort design. My investigations have shown three key areas where the case-cohort design affects estimation and post-estimation procedures, which I now discuss. In Section 8.3.1 I discuss the effects of case-cohort characteristics such as non-case to case ratio and subcohort size, with particular regard to choice of weighting method and impact of entry type on risk set size and variability. In Section 8.3.2 I discuss use of weighting to account for the over-estimation of cases in the case-cohort design. Finally, in Section 8.3.3 I discuss the importance of careful choice of variance estimation methods.

8.3.1 Effects of Case Cohort Characteristics

In Chapter 3, I note that methodological papers regarding the case-cohort design tend to design simulation studies so as to highlight changes in performance from the full cohort and often consider scenarios that are not reflective of the case-cohort design in practice. Further, in the literature, results and conclusions are often described with reference to full cohort characteristics. Also in Chapter 3, I theorized that case-cohort characteristics and influence of entry type may be more relevant to the behaviour of estimation and post-estimation methods than a simple description of the full cohort and the subcohort sampling fraction. Such characteristics include size and variability of risk sets, number of cases in the sample, subcohort size, subcohort sampling fraction and non-case to case ratio.

The results from previous chapters provide evidence supporting this approach. To illustrate this, I first in Section 8.3.1.1 consider variability in the full cohort for $R_{(j)}$, the observations at risk at time $t(j)$, and then in Section 8.3.1.2 discuss the effects of case-cohort characteristics and weighting method. Then, in Section 8.3.1.3 I discuss the results of the simulation studies in this context.

8.3.1.1 Risk Sets in the Full Cohort

Estimation and post-estimation procedures rely upon a number of quantities derived from risk sets. For example, estimation of coefficients and cumulative baseline hazards relies upon quantities calculated for each risk set - the denominator in both cases being $\sum_{i \in R_{(j)}} exp(\beta^T Z_i)$. Schoenfeld residuals are calculated for each individual risk set, defined as the difference between the value of the covariate and its mean conditioned upon the risk set at the failure time of that observation.

Smaller risk sets will see the quantities mentioned above become more variable. Further, more variable covariates Z will also lead to greater variation. Recall that under staggered entry, one would expect for risk sets to be smaller than for a similar dataset analysed under fixed entry. Further, the sizes of risk sets across the entire dataset would be more variable also.

8.3.1.2 Risk Sets in the Case Cohort

Smaller sampling fractions will lead to the subcohort risk set $R*_{(j)}$ being more variable, relative to the full cohort. Note further that due to random chance, individual risk sets may see sampling fractions quite discordant from the overall sampling fraction.

In Prentice weighting, additional variation relative to the full cohort is present for both cases and non-cases, as non-subcohort cases are included only at their failure time. Note further that, as the proportion of non-subcohort cases in the case-cohort dataset increases, a greater proportion of risk sets will include a non-subcohort case. At smaller subcohort sizes, or where Z for cases is more discordant from Z for non-cases, this non-subcohort case will have greater influence on the quantities mentioned above. Further, Prentice weighting will lose more information from exclusion of non-subcohort cases from the risk sets.

By contrast, in IPW weighting, the cases are identical to the full cohort in all respects, and variance due to sampling occurs only for the non-cases. However, note that where the weights for non-cases are calculated from the overall sampling fraction, they may be discordant from the sampling fraction for the individual risk set $R*_{(j)}$.

8.3.1.3 Effect of Case-Cohort Characteristics on Case Cohort Estimation and Post-Estimation Procedures

Given Section 8.3.1.2, one might expect for IPW weighting to be generally superior to Prentice weighting. Further, one might expect results under fixed entry to be generally superior to results under staggered entry, where risk sets are smaller and more variable. However, this was not seen in the simulation studies. This can be reconciled by considering that 8.3.1.1 and 8.3.1.2 consider individual risk sets, and that a number of estimation and post-estimation procedures, to a greater or lesser degree, in some way aggregate estimates or quantities from individual risk sets to achieve their final results. For example, in estimation of β, model selection, and use of martingale residuals to detect inappropriate functional form, only minor differences were seen between Prentice weighting and IPW. Results for staggered entry and fixed entry were also very similar in the chapters on Martingale residuals and Model Selection. Note that the number of cases in the dataset not only increases the amount of information, it also increases the number of risk sets over which such aggregation takes place. The differences between bias and empirical standard error of estimates of β between entry types was most apparent with smaller numbers of cases.

However, where procedures are more sensitive to size and variability of individual risk sets, more profound differences between entry types and case-cohort weighting methods were seen. In detection of non-proportional hazards, when all covariates were allowed to interact with time simultaneously and a global test for their significance was performed, inappropriately high type 1 error rates were associated with

smaller numbers of cases, particularly in staggered entry. When p-values were considered separately for each covariate, Prentice weighting showed somewhat higher Type 1 error rate and lower power than IPW. Type 1 error rate was also increased in staggered entry compared to fixed entry.

Results for methods incorporating Schoenfeld residuals were particularly interesting. Recall that Schoenfeld residuals are calculated for each individual risk set. With these methods, entry type had a profound effect on performance. In fixed entry, both IPW and Prentice weighting showed similar Type 1 error to the full cohort for correlation of Schoenfeld residuals with time, and a slight elevation for the method of Grambsch & Therneau. In staggered entry, both IPW and Prentice showed highly inappropriate Type 1 error rate for both these methods in all circumstances. When the proportion of non-subcohort cases was low, IPW showed higher type 1 error than Prentice. Where the proportion of non-subcohort cases was high, IPW showed lower type 1 error than Prentice. It appears, therefore, that both interactions of covariates with time, and methods based on Schoenfeld residuals are more sensitive than other post-estimation methods to both the proportion of non-subcohort cases and the size and variability of the risk sets.

8.3.1.4 Conclusion

Overall, results from this thesis indicate that regard to case-cohort characteristics such as non-case to case ratio, subcohort size, and proportion of non-subcohort cases in the dataset, together with the effects of entry type on variability and size of risk sets, can be valuable. Validity and performance of estimation and post-estimation methods, and choice of weighting system, is better assessed by considering such characteristics than by a simple assessment of full-cohort characteristics and the subcohort sampling fraction.

8.3.2 Weighting in Post-Estimation

In the case-cohort design, cases are over-represented in the case-cohort sample. In this thesis, IPW weighting was used to account for this where estimation or imitation of a full-cohort quantity was desired, such as for mean-centering or smooths of martingale residuals against covariate values. In general, IPW Classic appeared appropriate for such aims.

Use of time-specific weights calculated for each individual risk set gave improvement in precision for estimation of cumulative baseline hazard, particularly early in analysis time where results from each individual risk set have more influence on

the overall estimate. The results from Chapter 5 further indicate that use of time-specific weights may be more valuable where quantities are calculated or estimated for specific risk sets such as for Schoenfeld residuals and scaled Schoenfeld residuals. However, use of time-specific weights is computationally intensive and in other circumstances may not result in gains in sufficient gains in performance to justify the increased computational time for analysis.

8.3.3 Estimation of Variance

It is well known that naïeve estimation of variance that does not account for case-cohort design gives inappropriately small results for variance of β. As discussed in 1.3.1.7, a number of methods for estimation of coefficient sampling variance that account for the covariance between score terms in the case-cohort design have been proposed.

Further, this thesis has shown that careful consideration is also required for estimation of variance for post-estimation methods in the case-cohort design. As described in Chapter 6, use of $dAIC$, $dBIC$ and dLR requires calculation of design effects using both naïeve and robust estimates of variance. In Chapter 5, there are indications that substitution of the average variance of the covariates for the weighted variance of the covariance at each failure time in the methods of Grambsch and Therneau (1994), is inappropriate in the case-cohort, even where it is valid in the full cohort.

8.4 Limitations

In this thesis I have not considered the effect of stratified Cox modelling on the estimation and post-estimation procedures investigated. In certain post-estimation procedures such as calculation of martingale residuals and Schoenfeld residuals, the stratum-specific case-cohort characteristics appear more relevant than the characteristics of the case-cohort sample as a whole. The degree of any such impact has not been assessed in this thesis. In the review of the case-cohort design in practice by Sharp et al. (2014), 9 of the 17 original cohorts used stratified sampling to select the subcohort, indicating that evaluation of these methods in stratified Cox modelling would be of value.

In Chapter 4, the lack of a valid method to quantify results of visual assessment of smooths over a large number of replicates in a simulation study made assessment of the behaviour of smooths of martingale residuals against covariate values difficult to assess. In Section 4.4 I discuss the unsuitability of correlation coefficients and

inspection of the mean value of the smooth over the replications at benchmark covariate values. The use of statistical assessment of non-linearity of weighted linear splines as a proxy for subjective visual assessment of weighted smooths also appears inadequate.

The general data-generating mechanism used throughout this thesis has its own limitations, as discussed in Section 2.8. Briefly, loss to follow-up was also not considered in this thesis. A degree of loss-to-follow-up is likely to occur in real-world studies, and may impact on the findings of this thesis. In general, loss-to-follow-up will result in smaller and more variable risk sets, the effect of which has been evident throughout this thesis. A Weibull model was used for simulation of survival times, with parameters modelled from the Cambridge Centre of the Interact dataset and λ scaled to provide consistency in survival times. The administrative censoring procedure and subcohort sampling procedure also aimed to provide consistency in case percentages and empirical sampling fraction. The scope of simulation studies in this thesis could hence be considered limited to datasets that are similar to the Cambridge Centre of the Interact dataset.

8.5 Future Work

Areas for future work in the case-cohort design are extensive. Ideally, one would seek to be able to apply to the case-cohort all the methods that might be used to analyse a full cohort. However, below I outline four key areas in which I believe further investigation would be particularly useful.

8.5.1 Application of Design Effects

The usefulness of dLR, $dBIC$, $dAIC$ and $pBIC$, each originally proposed for the field of complex survey sampling, highlights the value of considering complex survey sampling methods for application to the case cohort design. It is possible that the field of complex survey sampling has additional methods that can be applied to the case-cohort design. With the exception of $pBIC$, each of these methods makes use of a design effect, a measure of the loss of effectiveness by the study design as compared to a simple random sample. Use of design effects may allow for adaptation of full cohort methods to the case cohort design, particularly where such methods rely on estimates of variance.

For example, Lin et al. (1993) propose graphical measures and numerical tests for assessing model fit in the Cox model based on cumulative sums of martingale residuals.

They describe tests for the functional form of a covariate, the form of the link function, the validity of the proportional hazards assumption, and an omnibus test for model misspecification. These tests are based upon counting processes W, where martingale residuals are summed cumulatively over covariate values and/or time, and the limiting distribution of W is approached through Monte Carlo simulations, with W simulated by replacing various unknown quantities with their respective consistent estimates. The observed process w can be plotted against realizations from W for a graphical check of model adequacy. The numerical test is to derive 1000 realizations from W and to calculate the proportion of times that the maximum absolute value of w is less than the maximum absolute value of each realization, with a p-value of $\leq .05$ evidence for rejection of the null hypothesis of adequate model fit.

In exploratory work not presented in this thesis I had attempted to apply this test for functional form of a covariate to the case-cohort design. IPW weighting of the residuals provided similar graphical "shape" to the full cohort w, however, the realizations from W were overly conservative, with low variance, and with even the correct model displaying extremely low p-values. The methods of Lin et al. (1993) for simulating realizations from W in the full cohort provide two possible sources of increased "spread" of the realisations from W. The first is $\mathcal{I}$, the observed information of the full cohort. The second is random draws from a standard normal distribution $G \sim (N(0,1)$ with Martingale residuals $M_i(t_{(j}))$, estimated in the formula by $N_i(t_{(j)}))G_i$ where $N_i(t_{(j)}))$ is the observed counting process. This approach is justified as the variance function of $M_i(t_{(j}))$ is $E(N_i(t_{(j)})))$. It is possible that incorporating design effects may be useful in modifying these methods for the case-cohort design.

8.5.2 Case-Cohort Scaled Schoenfeld Residuals

In Chapter 5, I found that the scaled Schoenfeld residuals test of Grambsch and Therneau (1994), as implemented in STATA, was inappropriate for use in detection of non-proportional hazards in the case cohort design. I theorized that the high type 1 error rates seen could be due to an effect noted by Winnett and Sasieni (2001), that the average variance of the covariates is a poor proxy for the weighted variance of the covariance at each failure time when the variance of the covariates changes substantially over each failure time. In Chapter 7, I noted that a successful implementation of the methods of Grambsch and Therneau (1994) would be much less computationally intensive than methods using interactions of covariates with time, and would allow for more practical assessment of potential violations of the assumption of proportional hazards. Investigation of the use of time-specific weighted

variances of the covariates in the methods of Grambsch and Therneau (1994) would hence be very valuable in the case cohort design.

8.5.3 Global Goodness of Fit

In addition to the omnibus test of Lin et al. (1993) above, a number of global goodness-of-fit statistics have been proposed for the Cox Proportional Hazards model in the full cohort (e.g. Schoenfeld (1980) Grønnesby and Borgan (1996)). The tests, in general form, compare the model to an alternative, which includes indicator variables for discrete partitions of the covariate space (May and Hosmer, 1998). Formal guidance on partitioning the covariate space is given by Parzen and Lipsitz (1999) which allows for detection of need for interactions or higher order powers of covariates in the model. Future work could include investigation of these methods in the case-cohort design.

8.5.4 Firm Guidelines on Case-Cohort Characteristics

In Section 8.3.1 above, I discuss how estimation and post-estimation methods are affected by various case-cohort characteristics. In my simulation studies I found various circumstances in which, for example, a particular number of cases was too small and a particular larger number of cases was adequate. For guidance of analysis, clear thresholds or more comprehensive guidelines for the case-cohort characteristics required for a particular method to be valid would be of use. In particular, investigation of the effects of stratified Cox analysis, where simple consideration of the characteristics of the whole case-cohort sample may be insufficient, would be valuable.

8.6 Conclusion

In Chapter 1 I said that to fully exploit the case-cohort design, we must look beyond parameter estimation. With this thesis, I sought to investigate existing methods, adapt existing methods, and devise new methods for post-estimation in the case-cohort design.

I have shown that IPW methods of estimation of β and cumulative baseline hazard are appropriate in the case-cohort design. I have shown that weighted smooths of martingale residuals may be informative in assessment of appropriate functional forms of continuous covariates in the case cohort design, but that loss of power from the full cohort may be substantial, and in excess of that seen for other case-cohort estimation and post-estimation methods. I have shown that inclusion of interactions

of covariates with time as a method of assessment of violation of the proportional hazards assumption is appropriate in the case-cohort design, when risk set sizes and number of cases are not overly small. In this work, approximately 250 cases sufficed for reasonable Type 1 error rate, but further work to define thresholds to guide analysts would be of value. Further, I show that methods incorporating Schoenfeld residuals cannot be directly applied to the case-cohort design when risk sets are small and/or variable, as may arise in staggered entry. I have shown that dLR, $pBIC$, and $dAIC$ are valuable methods of variable and model selection in the case-cohort design, especially where data is sparse, and that Wald tests and $dBIC$ may be inappropriate where data is sparse. Finally, I have shown how use of these methods allowed for a more comprehensive investigation and analysis in a real-world case-cohort dataset.

I hope, therefore, that this dissertation makes steps towards allowing for greater exploitation of the case-cohort design in practice, and more comprehensive analysis of the valuable datasets that are both extant and will be available in the future.

Appendix A

Additional Results for Estimation

A.1 Additional Results for Estimation of β

In general, MCSE bounds for Type 1 error encompassed a nominal 5%, with exceptions generally for staggered entry and the 3% sampling fraction. Where Type 1 error was underestimated, upper bounds did not fall below $\sim$ 4.75%, point estimates did not fall below $\sim$ 3.4% and lower bounds did not fall below $\sim$ 2.25 %. Where Type 1 error was overestimated, upper bounds did not exceed $\sim$ 9%, point estimates did not exceed $\sim$ 7.5%, and lower bounds did exceed $\sim$ 6%.

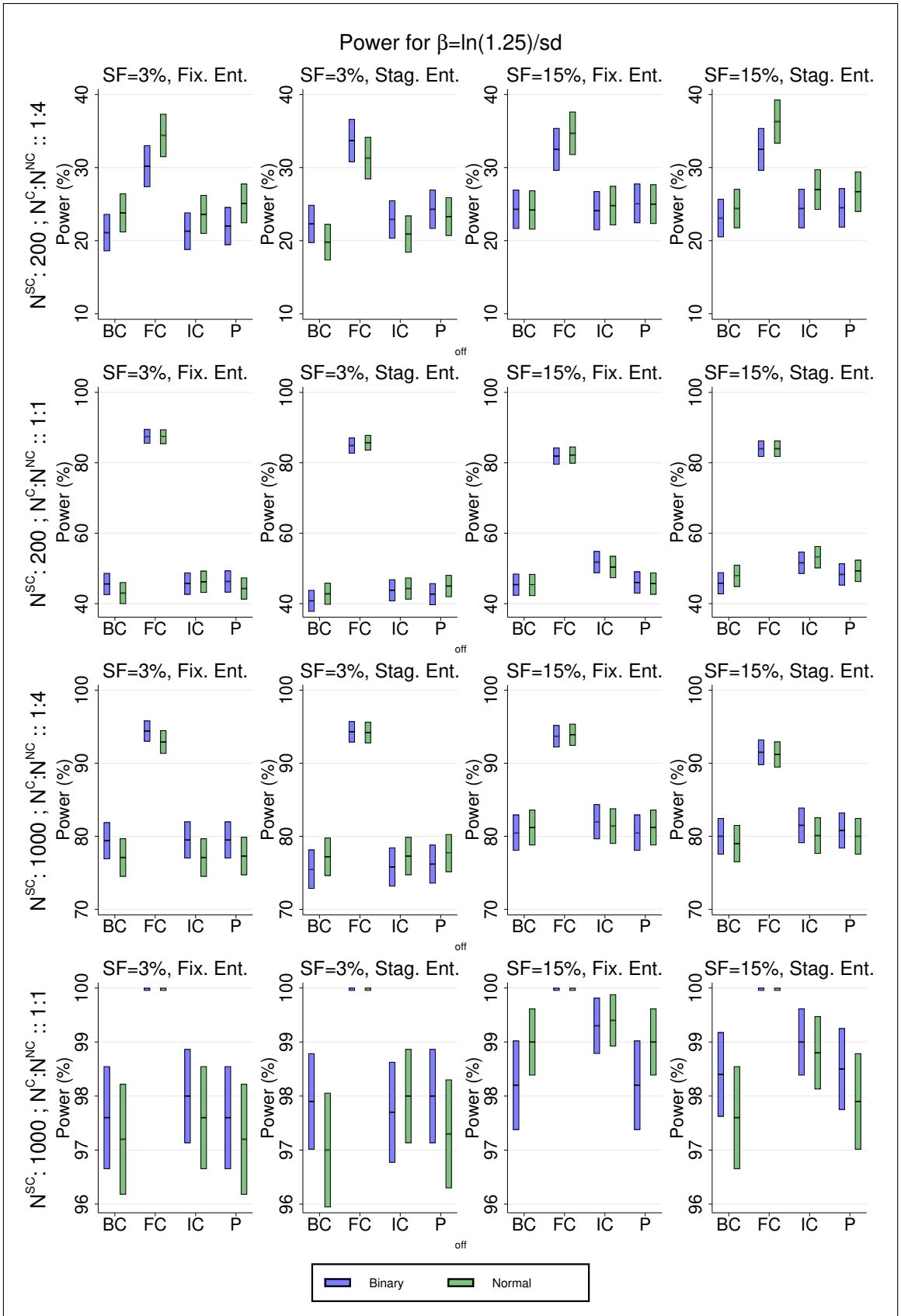
MCSE bounds for proportional error in the model-based standard error were within 5% of 0 except for certain combinations at a subcohort size of 200 where they were below -5%. Upper MCSE bounds did not fall below $\sim$ -10%, point estimates did not fall below $\sim$ -15%, and lower MCSE bounds did not fall below $\sim$ -20% All case-cohort weighting methods showed similar (overlapping MCSE bounds) proportional error in the model-based standard error.

MCSE bounds for coverage tended to fall below 95% at a subcohort size of 200, predominantly for the normal covariate with $\beta = \ln(2)$. Bounds fell below 92% only for staggered entry with a 3% sampling fraction, case to non-case ratio of 1, and the normal covariate with $\beta = \ln(2)$ (point estimates $\sim$ 89%). All weighting methods showed similar (overlapping MCSE bounds) coverage. Power for $\beta = \ln(1.25)/\text{SD}$ is presented in Figure A.1. IPW Classic demonstrates a small improvement over Prentice at subcohort size 200, sampling fraction 15% and case to non-case ratio 1:1, but this improvement does not cause MCSE bounds for these weighting methods to fail to overlap. Power for $\beta = \ln(2)/\text{SD}$ exceeds 94% for all weighting methods at a subcohort size of 200 and case to non-case ratio 1:4, and is 100% for all other combinations.

Figure A.1: Point Estimates & MCSE Bounds for Power for $\beta = \ln(1.25)/\text{SD}$

(a) BC = Barlow Classic; FC= Full Cohort; IC = IPW Classic; P = Prentice

(b) Shaded bars indicate MCSE bounds



A.2 Additional Information on Results for Estimation of $H_0(t)$

Table A.1: Minimum and Maximum Values for True $H_0(t)$

HR/SD	N^{SC}	α^{NC}	Ratio	Staggered		Fixed	
				Min	Max	Min	Max
1.1	200	3	1	0.0020	0.1154	0.0003	0.0282
			4	0.0002	0.0234	0.0002	0.0082
		15	1	0.0086	0.5322	0.0014	0.1415
			4	0.0083	0.1206	0.0005	0.0355
	1000	3	1	0.0030	0.1141	0.0003	0.0296
			4	0.0007	0.0273	0.0001	0.0072
		15	1	0.0169	0.5544	0.0012	0.1382
			4	0.0027	0.1401	0.0047	0.0379
2	200	3	1	0.0014	0.0714	0.0002	0.0191
			4	0.0001	0.0150	0.0000	0.0051
		15	1	0.0056	0.3463	0.0009	0.0947
			4	0.0006	0.0770	0.0003	0.0233
	1000	3	1	0.0027	0.0784	0.0002	0.0198
			4	0.0002	0.0192	0.0001	0.0048
		15	1	0.0099	0.3578	0.0008	0.0972
			4	0.0014	0.0888	0.0002	0.0239

In staggered entry, the difference in bias between Prentice Time and Prentice Classic does not exceed 2.5% of the true value of $H_0(t)$ in the first 10 reference times and does not exceed 1.5% of true $H_0(t)$ in the remainder of analysis time. MCSE bounds fail to overlap only at reference time 88+ in subcohort size 200, sampling fraction 3%, case to non-case ratio 1:1 and $\beta = \ln(1.1)/\text{SD}$. In the first 10 reference times, difference in bias between IPW Time and IPW Classic ranges from -1.1% of true $H_0(t)$ to +2.5% of true $H_0(t)$ and from -1.5% of true $H_0(t)$ to +1.4% of true $H_0(t)$ in the remainder of analysis time. MCSE bounds fail to overlap in the first

10 reference times where $\beta = \ln(1.1)/\text{SD}$.

In staggered entry, difference in empirical standard error between Time and Classic variants, for Prentice and IPW, respectively, does not exceed 5% of true $H_0(t)$ and 5.6% of true $H_0(t)$ in the first 10 reference times and does not exceed 3% of true $H_0(t)$ and 3.9% of true $H_0(t)$ in the remainder of analysis time.

In staggered entry, MCSE bounds for bias of Prentice Classic fail to overlap with those of the full cohort at $\beta = \ln(2)/\text{SD}$, sampling fraction 3%, subcohort size 200, and case to non-case ratio 1:1. MCSE bounds for bias of Prentice Time fail to overlap with those of the full cohort at sampling fraction 3% and case to non-case ratio 1:1; subcohort size 200, sampling fraction 15%, and case to non-case ratio 1:1; and subcohort size 200, sampling fraction 3%, case to non-case ratio 1:4, and $\beta = \ln(2)/\text{SD}$;

In staggered entry, MCSE bounds for empirical standard error of IPW Classic and Prentice Classic fail to overlap with those of the full cohort at $\beta = \ln(1.1)/\text{SD}$ when case to non-case ratio is 1:1. MCSE bounds for empirical standard error of IPW Classic fail to overlap with those of the full cohort at $\beta = \ln(2)/\text{SD}$ and case to non-case ratio 1:4. MCSE bounds for empirical standard error of all case-cohort estimators fail to overlap with those of the full cohort at $\beta = \ln(2)/\text{SD}$ and case to non-case ratio 1:1; and at $\beta = \ln(2)/\text{SD}$, case to non-case ratio 1:4, subcohort size 1000 and sampling fraction 3%.

Appendix B

Appendix for Chapter 7

Table B.1: InterAct Study: Assessment of Improved Model Fit with Restricted Cubic Splines (WC as Obesity Measure)

WC		dAIC Difference from Linear Form	pBIC Difference from Linear Form	dBIC	Wald	dLR
Male	BMI	-279.89	-257.83	222.97	<0.001	<0.001
	Alcohol	-3.42	2.63	-0.11	0.071	0.076
Female	BMI	-588.74	-569.89	500.87	<0.001	<0.001
	Alcohol	-27.15	-19.99	21.47	<0.001	<0.001

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

Table B.2: InterAct Study: Assessment of Improved Model Fit with Restricted Cubic Splines (WHtR as Obesity Measure)

		dAIC Difference from Linear Form	pBIC Difference from Linear Form	dBIC	Wald	dLR
Male	BMI	-469.44	-438.53	359.93	<0.001	<0.001
	Alcohol	-4.76	1.35	1.23	0.057	0.061
Female	BMI	-612.18	-596.24	526.72	<0.001	<0.001
	Alcohol	-12.70	-16.73	18.24	0.002	0.001

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, BMI

Table B.3: InterAct Study:Wald Tests for Interaction of Covariates with Rank Time (WC as Obesity Measure)

	PA	BMI	Alcohol	Calories	School	Smoke
Men	0.089	<0.001	0.644	0.257	0.862	0.702
Women	0.584	<0.001	0.553	0.303	0.596	0.246

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, Obesity Measure

Table B.4: InterAct Study:Wald Tests for Interaction of Covariates with Rank Time (WHtR as Obesity Measure)

	PA	BMI	Alcohol	Calories	School	Smoke
Men	0.181	<0.001	0.758	0.174	0.496	0.867
Women	0.483	<0.001	0.524	0.459	0.496	0.273

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, Obesity Measure

Table B.5: Hazard Ratios for Physical Activity - Comparison of Model Steps (WC as Obesity Measure)

		Interact 2012 Model	Stratified Cox Model	Adjusted for Functional Form	Adjusted for Non-Proportional Hazards
Men	HR	0.93	0.94	0.95	0.95
	95% CI	0.86, 1.00	0.89, 0.98	0.91, 0.99	0.91, 0.99
Women	HR	0.93	0.94	0.95	0.95
	95% CI	0.89, 0.99	0.89, 0.99	0.91, 1.00	0.91, 100

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WC

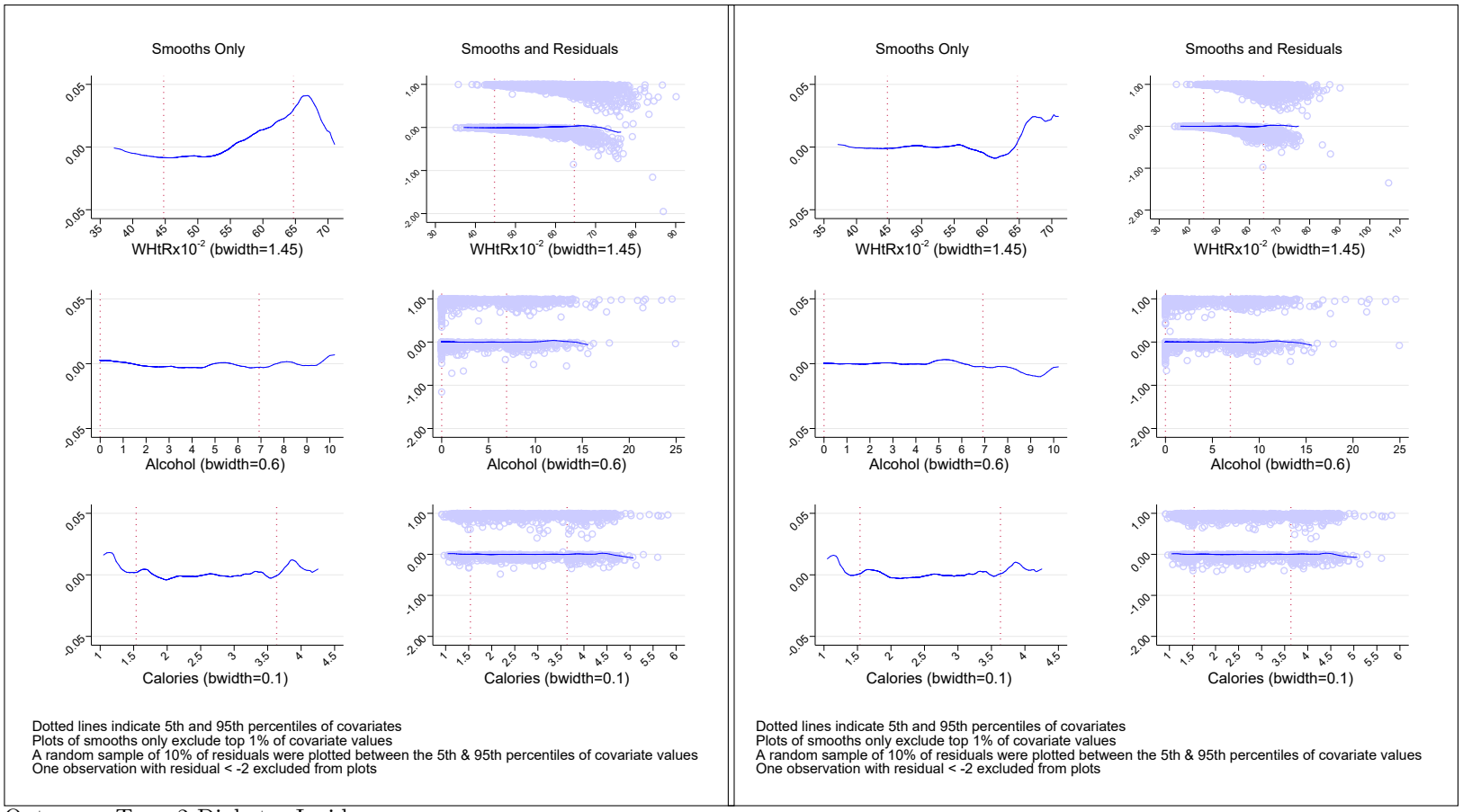
Table B.6: Hazard Ratios for Physical Activity - Comparison of Model Steps (WHtR as Obesity Measure)

WHtR		Interact 2012 Model	Stratified Cox Model	Adjusted for Functional Form	Adjusted for Non-Proportional Hazards
Men	HR	n/a	0.93	0.94	0.94
	95% CI	n/a	0.88, 0.97	0.90, 0.98	0.90, 0.99
Women	HR	n/a	0.95	0.96	0.95
	95% CI	n/a	0.90, 1.00	0.91, 1.00	0.91, 1.00

Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WHtR

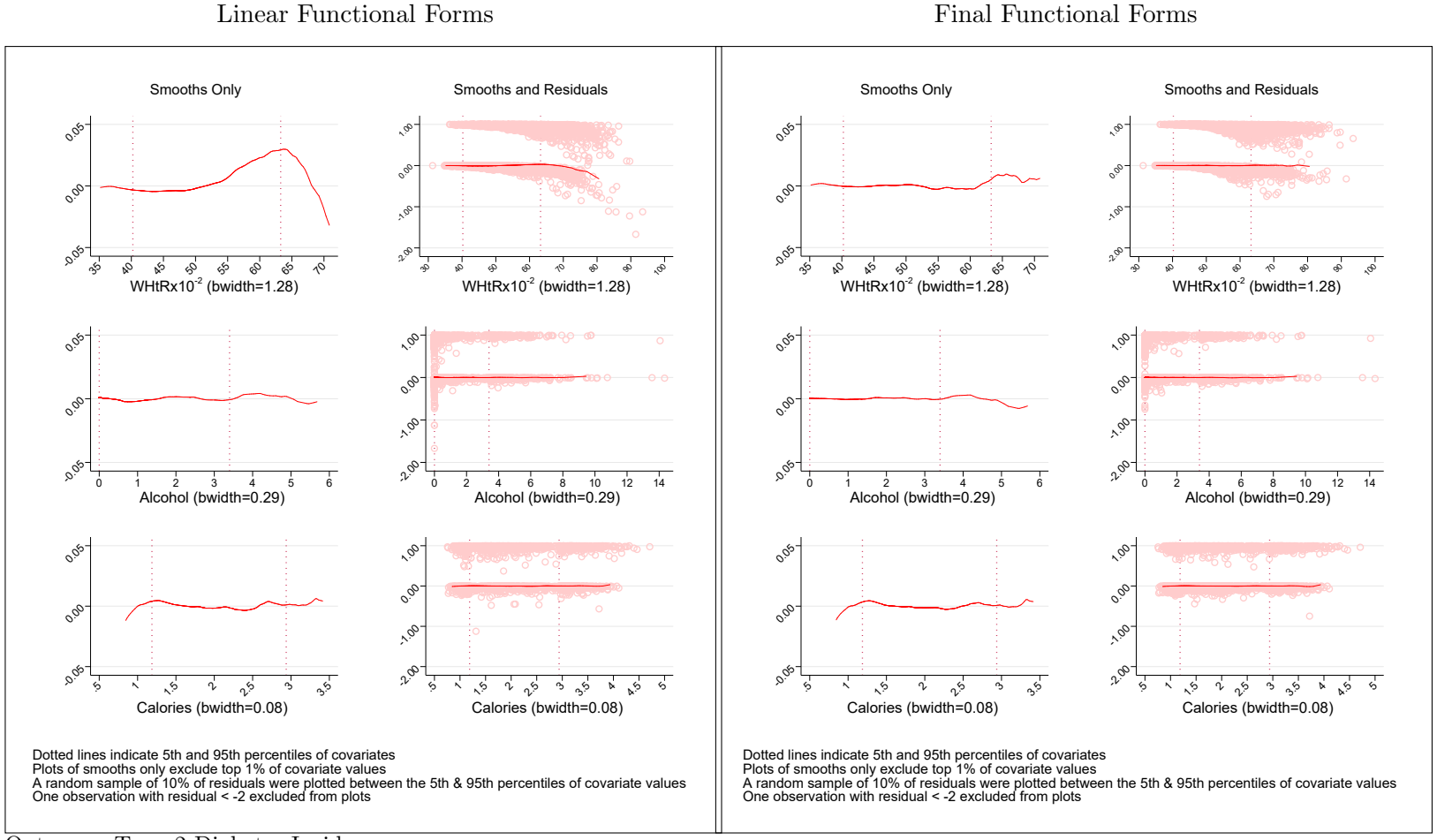
Figure B.1: InterAct Study: Smooths of Martingale Residuals in Men (WHtR)



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WHtR

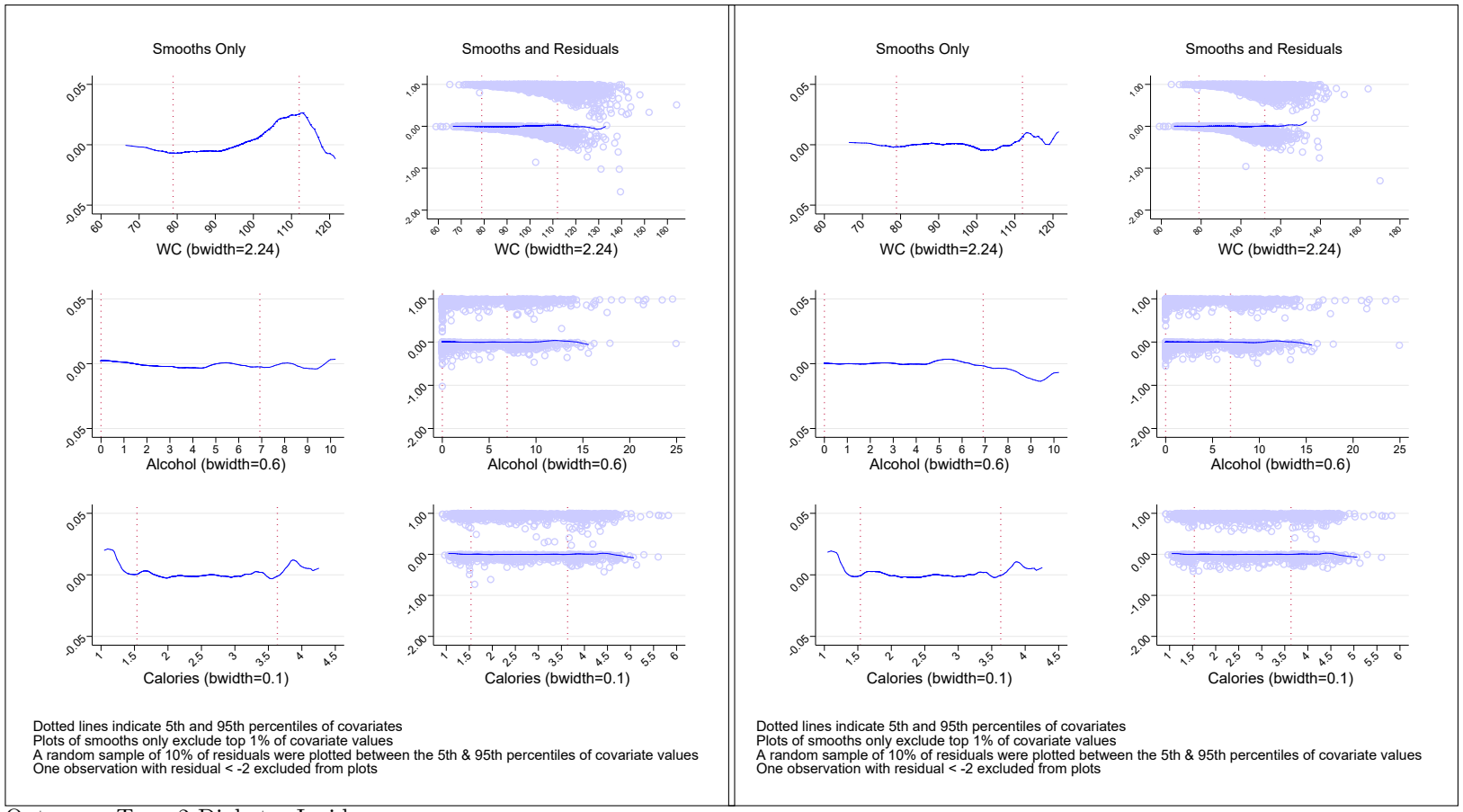
Figure B.2: InterAct Study: Smooths of Martingale Residuals in Women (WHtR)



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WHtR

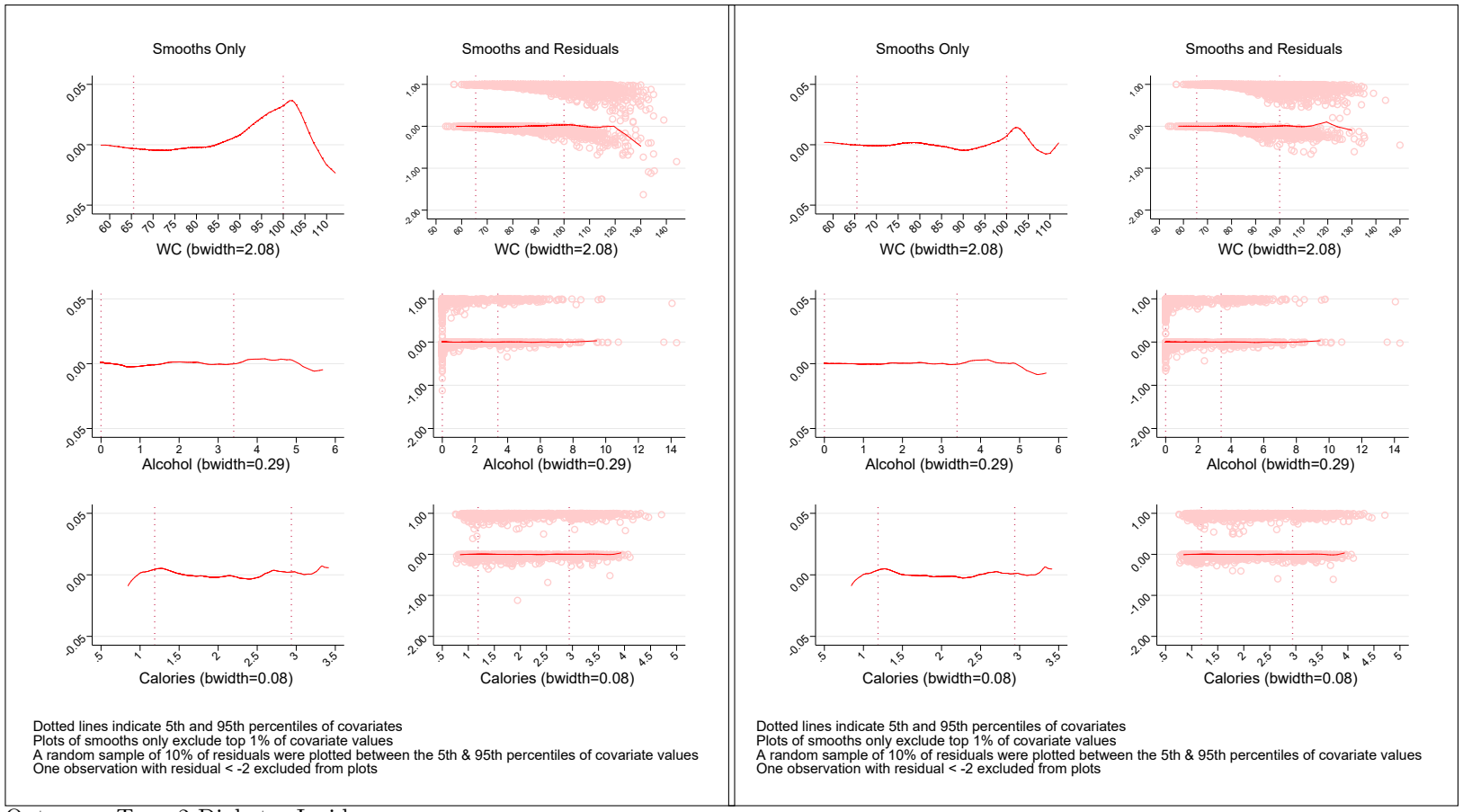
Figure B.3: InterAct Study: Smooths of Martingale Residuals in Men (WC



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WC

Figure B.4: InterAct Study: Smooths of Martingale Residuals in Women (WC)



Outcome: Type 2 Diabetes Incidence

Model Includes Physical Activity, Calories, Smoking, School, Alcohol, WC

Appendix C

Stata Comments & Sample Code

C.1 Introduction

In this appendix, STATA (Version 15.1) sample code for the methods described in this thesis is presented for IPW and Prentice weighting, together with additional comments on the implementation of the methods in STATA.

C.2 Data Setup

Package `carryforward`, available from `SSC` is required. The sample code given in this appendix is for a dataset with the following variables, scalars, and locals. Sample code to create an example dataset in this format is given below.

- The variable `id` records subject id.
- The variable `stime` records time of event or censoring.
- The variable `etime` records time of entrance to study.
- The variable `case` records case or non-case status with 0=non-case, 1=case.
- The variable `_subco` records subcohort status with 0 = subcohort case, 1 = subcohort non-case, 2=non-subcohort case.
- The variable `W_IPW` records the IPW weights
- The variable `W_Pren` records the weights for cumulative baseline hazards in Prentice weighting.
- The variables `X1`, `X2`, `X3` record values for predictor variables.
- The local `covlist` lists the covariates included in the model.

```
// ~~~Set up sample case-cohort dataset~~~
clear
set obs 5000
gen id = _n
gen X1 = rnormal(0,1)
gen X2 = rnormal(0,1)
gen X3 = rnormal(0,1)
gen etime = rnormal(10,1)
local beta 0.5 //define coefficient
local shape 3 //Weibull Parameters
local lambda 5.0e-5
//generate survival times
gen u1 = runiform(0,1)
gen stime = (etime^`shape' - log(u1)/(`lambda'*exp(X1*`beta')))^(1/`
    ↪ shape')
//admin. censoring for desired case percentage CP
local CP =0.05
gen ftime = stime - etime
sort ftime
gen case = cond(_n <= _N*`CP', 1, 0)
replace ftime = ftime[_n-1] if case ==0
replace stime = etime+ftime
count if case ==0
scalar FC_NC_N = r(N)
//Sample dataset at chosen level SF
local SF = 0.05
generate random = runiform()
sort random
generate _subco = cond(case ==1 & _n <= _N*`SF', 1, cond(case ==1 &
    ↪ _n > _N*`SF', 2, cond(case !=1 & _n <= _N*`SF', 0, . )))
drop if _subco ==.
//Create Weights for CBH in Prentice and for IPW Classic
gen W_Pren = 1/`SF'
count if _subco ==0
scalar SC_NC_N = r(N)
gen W_IPW = cond(case ==1, 1, FC_NC_N/SC_NC_N)
drop u1 random
save sampledata, replace
```

C.3 Estimation and Prediction

C.3.1 IPW

Implementation of IPW weighting for estimation of coefficients and estimation of cumulative baseline hazard is straightforward. Once weights have been calculated and included in the stset command, robust standard errors are calculated automatically and further adjustments are unnecessary.

Prediction of the linear predictor, relative hazard, Cox-Snell residuals, Martingale residuals and Schoenfeld residuals is likewise straightforward, using STATA's inbuilt predict command.

```
//~~~ Estimation & Predictions - IPW~~~
use sampledata, clear
local covlist X1
//set as survival time data
stset stime [pw=W_IPW], failure(case) id(id) enter(etime)
stcox ‘covlist’ //fit model
predict xb, xb //predict linear predictor
predict hr, hr //predict relative hazard
predict h0t, basehc //predict baseline hazard contribution
predict CBH, basechazard //predict cumulative baseline hazard
predict mg, mg //predict martingale residuals
save IPW_Predictions, replace
```

C.3.2 Prentice

Implementation of Prentice weighting for estimation of coefficients and estimation of cumulative baseline hazard requires a number of adjustments. For estimation of coefficients, an adjustment must be made to the entry time of non-subcohort cases and robust standard errors must be specified in the `stcox` command. Prediction of the linear predictor and relative hazard are straightforward. Baseline hazard contribution, cumulative baseline hazard, and martingale residuals must be calculated manually.

Note that in stratified models, the sample code provided here for calculation of cumulative baseline hazard and martingale residuals must be run separately for each stratum.

```
//~~~ Estimation & Predictions - Prentice~~~
use sampledata, clear

stset stime, failure(case) id(id) enter(etime)
replace _t0 = _t-.01 if _subco ==2 //adjust entry for non-subcohort
    ↪ cases
stcox ‘covlist’, robust //fit model with robust variance estimate
predict xb, xb //predict linear predictor
predict hr, hr //predict relative hazard
//manually calculate h0t
 sts gen nd_j = d // create variable recording # of failures
 //stsplit at failure times with variable indicating risk sets
 stsplit, at(failures) riskset(risk)
 //calculate sum of weighted relative hazard for each failure
 gen w_hr = hr*W_Pren
 bysort risk: egen denom = sum(w_hr)
 //account for tied failures, missing data
 bysort stime case: gen dup = cond(_N==1,0,_n) if case ==1
 replace denom = . if case !=1 | xb ==.
 replace denom = denom/nd_j
 gen temp_h0t1 =1/denom
 gen temp_h0t2 = 1/denom if dup <=1
//manually calculate CBH
sort stime dup
gen temp_CBH = sum(temp_h0t2) if dup<=1
bysort id: egen h0t = max(temp_h0t1)
bysort id: egen CBH = max(temp_CBH)
//consolidate to single record per subject
drop nd_j risk w_hr denom dup temp_h0t1 temp_h0t2 temp_CBH
stjoin
sort stime case
carryforward CBH, replace
replace CBH = . if xb ==.
```

```
//Calculate Martingale Residuals
//calculate CBH at etime
expand 2, gen(etemp)
replace stime = etime if etemp==1
gen temp_eCBH = CBH if etemp ==0 & case==1
gsort stime -case
carryforward temp_eCBH, replace
replace temp_eCBH = 0 if temp_eCBH ==.
bysort id: gen temp_eCBH2 = temp_eCBH if etemp==1
bysort id: egen eCBH = max(temp_eCBH2)
drop if etemp==1
drop etemp temp_eCBH temp_eCBH2
gen mg = case-hr*(CBH-eCBH)

save Pren_Predictions, replace
```

C.4 Functional Form

As described previously, martingale residuals can be used for detection of inappropriate functional form of a covariate. Inspection of a smooth of the martingale residuals against the functional form should be approximately linear with slope 0 when the functional form is appropriate. Once martingale residuals have been calculated using the above methods, the choice of weighting system has no impact on the assessment of covariate functional form. Hence, only IPW weighted smooths are presented here. Note that STATA will not accept sampling weights in the local polynomial smooth command (lpoly). Frequency-style weights can be constructed as a multiple of IPW weights rounded to the nearest integer for use in lpoly.

```
//~~~Smooths Of Martingale Residuals Against Covariate Values~~~
use IPW_Predictions, clear
//generate frequency-style IPW weights
gen fIPW = round(W_IPW*10, 1)
//plot lpoly smooth of martingale residuals against X1
lpoly mg X1 [fweight=fIPW]
```

C.5 Detection of Non-Proportional Hazards

Inclusion of a time-varying covariate in the model is straightforward for all weighting systems. Following specification of the survival time data in the `stset` command and adjustment of entry time for subcohort non-cases in Prentice weighting, case-cohort methods do not differ from full cohort methods, except that, as for estimation of coefficients, one should ensure that robust variance is specified in the stcox command for Prentice weighting.

```
//~~~Test for NPH by Inclusion of Interactions with Time~~~
//IPW
use sampledata, clear
stset stime [pw=W_IPW], failure(case) id(id) enter(etime)
stcox ‘covlist’, tvc(‘covlist’)

//Prentice
stset stime, failure(case) id(id) enter(etime)
replace _t0 = _t-.01 if _subco ==2
stcox ‘covlist’, tvc(‘covlist’) robust // note robust option
```

C.6 Model Comparison and Variable Selection

The following section considers the use of the robust Wald test, and case-cohort modifications for the Likelihood Ratio Test, AIC and BIC.

Robust Wald tests are easily implemented in STATA. Recall that the robust option must be specified in the estimation command for Prentice weighting.

For likelihood-based methods, use of sampling weights in the `stset` command provides erroneous values for the pseudo-partial-log likelihoods. Hence, with IPW, the pseudo-partial-log likelihoods must be obtained by using the `stset` command with `iweights`.

All likelihood-based methods presented here require estimates of both the robust and model-based variance-covariance matrixes. For IPW weighting this requires estimation after using each of `iweights` and `pweights` in the `stset` command. For Prentice weighting it requires presence and absence of specification of robust variance in the `stcox` command.

The sample code below will provide results for a Wald test and a dLR test for comparison of the model M_1 nested within M_M, and the values of $dBIC$, $pBIC$, and $dAIC$ for both model M_1 and model M_M.

```
//~~~Model Selection~~~
local vars_M "X1 X2 X3"
local vars_1 "X1"
local vars_1b "X2 X3" //variables in M_M not included in M_1
local k_M : word count ‘vars_M’
local k_1 : word count ‘vars_1’
local k_1b : word count ‘vars_1b’

use sampledata, clear

//Extraction of required values - IPW
// First, iweights are used to extract the pseuodopartial log
   ↪ likelihoods for M_1 and M_M, the naieve variance matrices, and
   ↪  the naieve Wald test statistic (for dBIC)
stset stime [iw=W_IPW], failure(case) id(id) enter(etime)
stcox ‘vars_M’
local pLL_M=e(ll)
mat V_n_M = e(V)
test ‘vars_1b’
local WaldBIC_1 = r(chi2)
stcox ‘vars_1’
local pLL_1=e(ll)
mat V_n_1 = e(V)
// Next, pweights are used to extract the robust variance matrices
   ↪ and perform robust Wald test
stset stime [pw=W_IPW], failure(case) id(id) enter(etime)
stcox ‘vars_M’
mat V_r_M = e(V)
test ‘vars_1b’
local Wald_1 = r(chi2)
local Waldp_1 = r(p)
stcox ‘vars_1’
mat V_r_1 = e(V)
```

```
//Extraction of required values - Prentice
stset stime, failure(case) id(id) enter(etime)
replace _t0 = _t-.01 if _subco ==2
// First, extract the pseuodopartial log likelihoods for M_1 and M_M,
   ↪  the naieve variance variance matrices and the naieve Wald
   ↪ test statistic (for dBIC)
stcox ‘vars_M’
local pLL_M=e(ll)
mat V_n_M = e(V)
test ‘vars_1b’
local Wald_1 = r(chi2)
local Waldp_1 = r(p)
stcox ‘vars_1’
local pLL_1=e(ll)
mat V_n_1 = e(V)
// Next, extract the robust variance matrices, and perform robust
   ↪ Wald test
stcox ‘vars_M’, robust
mat V_r_M = e(V)
test ‘vars_M_1b’
local Wald_1 = r(chi2)
stcox ‘vars_1’, robust
mat V_r_1 = e(V)

//~~~Calculate pBIC~~~
count if case ==1
local pBIC_M = 2*‘pLL_M’ + ‘k_M’*ln(r(N))
local pBIC_1 = 2*‘pLL_1’ + ‘k_1’*ln(r(N))

//~~~Calculate dAIC~~~~
//calculate dAIC design effect matrices
mat I_M = inv(V_n_M)
mat DE_M = I_M*V_r_M
local delta_M = trace(DE_M)
local dAIC_M = -2*‘pLL_M’+2*‘delta_M’
mat I_1 = inv(V_n_1)
mat DE_dAIC_1 = I_1*V_r_1
local delta_1 = trace(DE_dAIC_1)
local dAIC_1 = -2*‘pLL_1’+2*‘delta_1’
```

```
//~~~Calculate design effects matrix for dLR and dBIC~~~~
//create holding matrices
mat V_22 = J(‘k_1b’, ‘k_1b’, .)
mat I_22 = J(‘k_1b’, ‘k_1b’, .)
mat I_11 = J(‘k_1’, ‘k_1’, .)
mat I_12 = J(‘k_1’, ‘k_1b’, .)
mat I_21 = J(‘k_1b’, ‘k_1’, .)

//fill V_22 and I_22
local c1 = 1
foreach e1 of local vars_1b{
local c2 = 1
foreach e2 of local vars_1b{
mat V_22[‘c1’, ‘c2’] = V_r_M[rownumb(V_r_M,"‘e1’"),colnumb(V_r_M,"‘e2
   ↪ ’")]
mat I_22[‘c1’, ‘c2’] = I_M[rownumb(I_M,"‘e1’"),colnumb(I_M,"‘e2’")]
local c2 = ‘c2’+1
}
local c1 = ‘c1’+1
}

if ‘k_1’ !=0{   // account for when M_1 is null
//fill I_11 and I_12
local c1 = 1
foreach e1 of local vars_1{
local c2 = 1
foreach e2 of local vars_1{
mat I_11[‘c1’, ‘c2’] = I_M[rownumb(I_M,"‘e1’"),colnumb(I_M,"‘e2’")]
local c2 = ‘c2’+1
}
local c3 = 1
foreach e3 of local vars_1b{
mat I_12[‘c1’, ‘c3’] = I_M[rownumb(I_M,"‘e1’"),colnumb(I_M,"‘e3’")]
local c3 = ‘c3’+1
}
local c1 = ‘c1’+1
}

//fill I_21
```

```
local c1 = 1
foreach e1 of local vars_1b{
local c2 = 1
foreach e2 of local vars_1{
mat I_21[‘c1’, ‘c2’] = I_M[rownumb(I_M,"‘e1’"),colnumb(I_M,"‘e2’")]
    ↪ local c2 = ‘c2’+1
}
local c1 = ‘c1’+1
}
}

//calculate design effect matrix, accounting for case where M_1 is
    ↪ null
if ‘k_1’ ==0 mat Des_Eff = (I_22)*V_22
else mat Des_Eff = (I_22-I_21*inv(I_11)*I_12)*V_22

//calculate parameters for gamma approximation
matrix eigenvalues eig_DE im = Des_Eff
scalar q1 = 0
scalar q2 = 0
forvalues m = 1/‘k_1b’{
        scalar q1 = q1+eig_DE[1,‘m’]
        scalar q2 = q2+ 2*(eig_DE[1,‘m’])^2
}
local g_hat = (q1^2)/q2
local theta_hat = q2/q1

//calculate dLR
local dLR_1 = 2*(‘pLL_M’-‘pLL_1’)
local dLRp_1 = 1-gammap(‘g_hat’, ‘dLR_1’/‘theta_hat’)

//calculate geometric mean of eigenvalues
local dbar_1 =1
  forvalues m = 1/‘k_1b’{
    local dbar_1 = ‘dbar_1’*eig_DE[1,‘m’]
  }
local d_bar = ‘dbar_1’/‘k_1b’

//calculate dBIC
```

```
count if case ==1
local dBIC_1= `WaldBIC_1'-`k_1b'*ln(r(N)/`dbar_1')
local dBIC_M = 0

//~~~~Display Results~~~~
display "Wald chi2 = `Wald_1' p = `Waldp_1'"
display "dLR = `dLR_1' p = `dLRp_1'"
display "pBIC M_M = `pBIC_M' M_1 = `pBIC_1'"
display "dBIC M_M = `dBIC_M' M_1 = `dBIC_1'"
display "dAIC M_M = `dAIC_M' M_1 = `dAIC_1'"
```

Bibliography

Abdullah, M. B. (1990). On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 39(4):455–460.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723.

Austin, P. C. (2018). Statistical power to detect violation of the proportional hazards assumption when using the cox regression model. *J. Stat. Comput. Simul.*, 88(3):533–552.

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):1064–1072.

Barlow, W. E. (1997). Global measures of local influence for proportional hazards regression models. *Biometrics*, 53(3):1157–1162.

Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *J. Clin. Epidemiol.*, 52(12):1165–1172.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Stat. Med.*, 24(11):1713–1723.

Binder, D. A. (1992). Fitting cox's proportional hazards models from survey data. *Biometrika*, 79(1):139–147.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.*, 6(1):39–58.

Breslow, N. E. (1972). Discussion of the paper by d. r. cox. *J. R. Stat. Soc.*, (34):216–217.

Chastang, C. (1983). A SIMULATION STUDY IN a 2-COVARIATE SURVIVAL MODEL IMPORTANCE OF THE PROPORTIONAL HAZARD ASSUMPTION. In *Controlled Clinical Trials*, volume 4, pages 148–148. ELSEVIER SCIENCE INC 655 ....

Chen, K. (2001). Generalized case–cohort sampling. *J. R. Stat. Soc. Series B Stat. Methodol.*, 63(4):791–809.

Chen, K. (2004). Statistical estimation in the proportional hazards model with risk set sampling.

Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with cox's model. *Biometrika*, 86(4):755–764.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *Br. J. Cancer*, 89(2):232–238.

Cole, S. R., Chu, H., and Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am. J. Epidemiol.*, 179(2):252–260.

Cox, D. R. (1972). Regression models and Life-Tables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 34(2):187–220.

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

Croxford, R. (2016). Restricted cubic spline regression : A brief introduction.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., and Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Stat. Med.*, 32(9):1584–1618.

Engle, R. F. (1984). Chapter 13 wald, likelihood ratio, and lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, volume 2, pages 775–826. Elsevier.

Fleming, T. R. and Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Ganguli, B., Naskar, M., Malloy, E. J., and Eisen, E. A. (2015). Determination of the functional form of the relationship of covariates to the log hazard ratio in a cox model. *J. Appl. Stat.*, 42(5):1091–1105.

Ganna, A., Reilly, M., de Faire, U., Pedersen, N., Magnusson, P., and Ingelsson, E. (2012). Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *Am. J. Epidemiol.*, 175(7):715–724.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Grant, S., Chen, Y. Q., and May, S. (2014). Performance of goodness-of-fit tests for the cox proportional hazards model with time-varying covariates. *Lifetime Data Anal.*, 20(3):355–368.

Greenland, S. (1986). Adjustment of risk ratios in case-base studies (hybrid epidemiologic designs). *Stat. Med.*, 5(6):579–584.

Groenwold, R. H. H., Sterne, J. A. C., Lawlor, D. A., Moons, K. G. M., Hoes, A. W., and Tilling, K. (2016). Sensitivity analysis for the effects of multiple unmeasured confounders. *Ann. Epidemiol.*, 26(9):605–611.

Grønnesby, J. K. and Borgan, O. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.*, 2(4):315–328.

Hess, K. R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Stat. Med.*, 14(15):1707–1723.

InterAct Consortium, Ekelund, U., Palla, L., Brage, S., Franks, P. W., Peters, T., Balkau, B., Diaz, M. J. T., Huerta, J. M., Agnoli, C., Arriola, L., Ardanaz, E., Boeing, H., Clavel-Chapelon, F., Crowe, F., Fagherazzi, G., Groop, L., Føns Johnsen, N., Kaaks, R., Khaw, K. T., Key, T. J., de Lauzon-Guillain, B., May, A., Monninkhof, E., Navarro, C., Nilsson, P., Nautrup Østergaard, J., Norat, T., Overvad, K., Palli, D., Panico, S., Redondo, M. L., Ricceri, F., Rolandsson, O., Romaguera, D., Romieu, I., Sánchez Pérez, M. J., Slimani, N., Spijkerman, A., Teucher, B., Tjonneland, A., Travier, N., Tumino, R., Vos, W., Vigl, M., Sharp, S., Langeberg, C., Forouhi, N., Riboli, E., Feskens, E., and Wareham, N. J. (2012a). Physical activity reduces the risk of incident type 2 diabetes in general and in abdominally lean and obese men and women: the EPIC-InterAct study. *Diabetologia*, 55(7):1944–1952.

InterAct Consortium, Langenberg, C., Sharp, S., Forouhi, N. G., Franks, P. W., Schulze, M. B., Kerrison, N., Ekelund, U., Barroso, I., Panico, S., Tormo, M. J., Spranger, J., Griffin, S., van der Schouw, Y. T., Amiano, P., Ardanaz, E., Arriola, L., Balkau, B., Barricarte, A., Beulens, J. W. J., Boeing, H., Bueno-de Mesquita, H. B., Buijsse, B., Chirlaque Lopez, M. D., Clavel-Chapelon, F., Crowe, F. L., de Lauzon-Guillan, B., Deloukas, P., Dorronsoro, M., Drogan, D., Froguel, P., Gonzalez, C., Grioni, S., Groop, L., Groves, C., Hainaut, P., Halkjaer, J., Hallmans, G., Hansen, T., Huerta Castaño, J. M., Kaaks, R., Key, T. J., Khaw, K. T., Koulman, A., Mattiello, A., Navarro, C., Nilsson, P., Norat, T., Overvad, K., Palla, L., Palli, D., Pedersen, O., Peeters, P. H., Quirós, J. R., Ramachandran, A., Rodriguez-Suarez, L., Rolandsson, O., Romaguera, D., Romieu, I., Sacerdote, C., Sánchez, M. J., Sandbaek, A., Slimani, N., Sluijs, I., Spijkerman, A. M. W., Teucher, B., Tjonneland, A., Tumino, R., van der A, D. L., Verschuren, W. M. M., Tuomilehto, J., Feskens, E., McCarthy, M., Riboli, E., and Wareham, N. J. (2011). Design and cohort description of the InterAct project: an examination of the in-

teraction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC study. *Diabetologia*, 54(9):2272–2282.

InterAct Consortium, Langenberg, C., Sharp, S. J., Schulze, M. B., Rolandsson, O., Overvad, K., Forouhi, N. G., Spranger, J., Drogan, D., Huerta, J. M., Arriola, L., de Lauzon-Guillan, B., Tormo, M.-J., Ardanaz, E., Balkau, B., Beulens, J. W. J., Boeing, H., Bueno-de Mesquita, H. B., Clavel-Chapelon, F., Crowe, F. L., Franks, P. W., Gonzalez, C. A., Grioni, S., Halkjaer, J., Hallmans, G., Kaaks, R., Kerrison, N. D., Key, T. J., Khaw, K. T., Mattiello, A., Nilsson, P., Norat, T., Palla, L., Palli, D., Panico, S., Quirós, J. R., Romaguera, D., Romieu, I., Sacerdote, C., Sánchez, M.-J., Slimani, N., Sluijs, I., Spijkerman, A. M. W., Teucher, B., Tjonneland, A., Tumino, R., van der A, D. L., van der Schouw, Y. T., Feskens, E. J. M., Riboli, E., and Wareham, N. J. (2012b). Long-term risk of incident type 2 diabetes and measures of overall and regional obesity: the EPIC-InterAct case-cohort study. *PLoS Med.*, 9(6):e1001230.

InterAct Consortium, Scott, R. A., Langenberg, C., Sharp, S. J., Franks, P. W., Rolandsson, O., Drogan, D., van der Schouw, Y. T., Ekelund, U., Kerrison, N. D., Ardanaz, E., Arriola, L., Balkau, B., Barricarte, A., Barroso, I., Bendinelli, B., Beulens, J. W. J., Boeing, H., de Lauzon-Guillain, B., Deloukas, P., Fagherazzi, G., Gonzalez, C., Griffin, S. J., Groop, L. C., Halkjaer, J., Huerta, J. M., Kaaks, R., Khaw, K. T., Krogh, V., Nilsson, P. M., Norat, T., Overvad, K., Panico, S., Rodriguez-Suarez, L., Romaguera, D., Romieu, I., Sacerdote, C., Sánchez, M. J., Spijkerman, A. M. W., Teucher, B., Tjonneland, A., Tumino, R., van der A, D. L., Wark, P. A., McCarthy, M. I., Riboli, E., and Wareham, N. J. (2013). The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the EPIC-InterAct study. *Diabetologia*, 56(1):60–69.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Stat. Med.*, 7(1-2):149–160.

Kalbfleisch-Prentice (1980). *Statistical analysis of failure time data*. John Wiley and Sons.

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, 96(4):887–901.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53(282):457–481.

Keele, L. (2010). Testing for nonproportional hazards in cox models. *Political Analysis - POLIT ANAL*, 18:189–205.

Kim, R. S. (2014). A new comparison of nested case–control and case–cohort designs and methods. *Eur. J. Epidemiol.*, 30(3):197–207.

Kim, S., Cai, J., and Lu, W. (2013). More efficient estimators for case-cohort studies. *Biometrika*, 100(3):695–708.

Knott, C., Bell, S., and Britton, A. (2015). Alcohol consumption and the risk of type 2 diabetes: A systematic review and Dose-Response meta-analysis of more than 1.9 million individuals from 38 observational studies. *Diabetes Care*, 38(9):1804–1812.

Koppes, L. L. J., Dekker, J. M., Hendriks, H. F. J., Bouter, L. M., and Heine, R. J. (2005). Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. *Diabetes Care*, 28(3):719–725.

Kulathinal, S. and Arjas, E. (2006). Bayesian inference from case–cohort data with multiple end-points. *Scand. Stat. Theory Appl.*, 33(1):25–36.

Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of Relative-Risk estimation in Case-Cohort studies. *J. Am. Stat. Assoc.*, 99(467):832–844.

Kupper, L. L., McMichael, A. J., and Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *J. Am. Stat. Assoc.*, 70(351a):524–528.

Li, X.-H., Yu, F.-F., Zhou, Y.-H., and He, J. (2016). Association between alcohol consumption and the risk of incident type 2 diabetes: a systematic review and dose-response meta-analysis. *Am. J. Clin. Nutr.*, 103(3):818–829.

Lin, D. Y. (2000). On fitting cox's proportional hazards models to survey data. *Biometrika*, 87(1):37–47.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.*, 84(408):1074–1078.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

Lin, J., Zhang, D., and Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics*, 62(3):803–812.

Lumley, T. and Scott, A. (2013). Partial likelihood ratio tests for the cox model under complex sampling. *Stat. Med.*, 32(1):110–123.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *J Surv Stat Methodol*, 3(1):1–18.

May, S. and Hosmer, D. W. (1998). A simplified method of calculating an overall goodness-of-fit test for the cox proportional hazards model. *Lifetime Data Anal.*, 4(2):109–120.

Miettinen, O. (1982). Design options in epidemiologic research. an update. *Scand. J. Work Environ. Health*, 8 Suppl 1:7–14.

Mirzaei, M. and Khajeh, M. (2018). Comparison of anthropometric indices (body mass index, waist circumference, waist to hip ratio and waist to height ratio) in predicting risk of type II diabetes in the population of yazd, iran. *Diabetes Metab. Syndr.*, 12(5):677–682.

Moreau, T., O'quigley, J., and Mesbah, M. (1985). A global goodness-of-fit statistic for the proportional hazards model. *Journal of the Royal*.

Newcombe, P. J., Connolly, S., Seaman, S., Richardson, S., and Sharp, S. J. (2018). A two-step method for variable selection in the analysis of a case-cohort study. *Int. J. Epidemiol.*, 47(2):597–604.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A(1/2):175–240.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Stat. Med.*, 16(6):611–626.

Ni, A., Cai, J., and Zeng, D. (2016). Variable selection for case-cohort studies with failure time outcome. *Biometrika*.

Onland-Moret, N. C., van der A, D. L., van der Schouw, Y. T., Buschers, W., Elias, S. G., van Gils, C. H., Koerselman, J., Roest, M., Grobbee, D. E., and Peeters, P. H. M. (2007). Analysis of case-cohort data: a comparison of different methods. *J. Clin. Epidemiol.*, 60(4):350–355.

Park, S. and Hendry, D. J. (2015). Reassessing schoenfeld residual tests of proportional hazards in political science event history analyses. *Am. J. Pol. Sci.*, 59(4):1072–1087.

Parzen, M. and Lipsitz, S. R. (1999). A global goodness-of-fit statistic for cox regression models. *Biometrics*, 55(2):580–584.

Pawitan, Y. (2000). A reminder of the fallibility of the wald statistic: Likelihood explanation. *Am. Stat.*, 54(1):54–56.

Persson, I. and Kamis, H. (2005). Bias of the cox model hazard ratio. *J. Mod. Appl. Stat. Methods*, 4(1).

Petersen, L., Sørensen, T. I. A., and Andersen, P. K. (2003). Comparison of case-cohort estimators based on data on premature death of adult adoptees. *Stat. Med.*, 22(24):3795–3803.

Pettitt, A. N. and Bin Daud, I. (1989). Case-Weighted measures of influence for proportional hazards regression. 38(1):51.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Riboli, E., Hunt, K. J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U. R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., Palli, D., Bueno-De-Mesquita, H. B., Peeters, P. H. M., Lund, E., Engeset, D., González, C. A., Barricarte, A., Berglund, G., Hallmans, G., Day, N. E., Key, T. J., Kaaks, R., and Saracci, R. (2002). European prospective investigation into cancer and nutrition (EPIC): study populations and data collection. *Public Health Nutr.*, 5(6B):1113–1124.

Saarela, O. and Kulathinal, S. (2007). Conditional likelihood inference in a case-cohort design: an application to haplotype analysis. *Int. J. Biostat.*, 3(1):Article 1.

Samuelsen, S. O., Ånestad, H., and Skrondal, A. (2007). Stratified Case-Cohort analysis of general cohort sampling designs. *Scand. Stat. Theory Appl.*, 34(1):103–119.

Sanderson, J., Thompson, S. G., White, I. R., Aspelund, T., and Pennells, L. (2013). Derivation and assessment of risk prediction models using case-cohort data. *BMC Med. Res. Methodol.*, 13:113.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2(6):110–114.

Scheike, T. H. and Martinussen, T. (2004). Maximum likelihood estimation for cox's regression model under Case–Cohort sampling. *Scand. Stat. Theory Appl.*, 31(2):283–293.

Schemper, M. (1992). Cox analysis of survival data with Non-Proportional hazard functions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(4):455–465.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for Case-Cohort studies. *Ann. Stat.*, 16(1):64–81.

Sharp, S. J., Poulaliou, M., Thompson, S. G., White, I. R., and Wood, A. M. (2014). A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS One*, 9(6):e101176.

Son, Y. J., Kim, J., Park, H. J., Park, S. E., Park, C. Y., Lee, W. Y., Oh, K. W., Park, S. W., and Rhee, E. J. (2016). Association of Waist-Height ratio with diabetes risk: A 4-year longitudinal retrospective study. *Endocrinol Metab (Seoul)*, 31(1):127–133.

Song, H. H. and Lee, S. (2000). Comparison of goodness of fit tests for the cox proportional hazards model. *Communications in Statistics - Simulation and Computation*, 29(1):187–206.

Su, Y., Ma, Y., Rao, W., Yang, G., Wang, S., Fu, Y., Liu, Y., Zhang, Y., You, Y., Yu, Y., and Kou, C. (2016). Association between body mass index and diabetes in northeastern china: Based on Dose-Response analyses using restricted cubic spline functions. *Asia. Pac. J. Public Health*, 28(6):486–497.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

Wacholder, S., Gail, M. H., Pee, D., and Brookmeyer, R. (1989). Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*, 76(1):117–123.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, 54(3):426–482.

White, I. R. (2010). simsum: Analyses of simulation studies including monte carlo error. *Stata J.*, 10(3):369–385.

WHO (2008). Waist circumference & Waist-Hip ratio. Technical report.

Winnett, A. and Sasieni, P. (2001). A note on scaled schoenfeld residuals for the proportional hazards model. *Biometrika*, 88(2):565–571.

Xu, C., Chen, J., and Mantell, H. (2013). Pseudo-likelihood-based bayesian information criterion for variable selection in survey data. *Surv. Methodol.*, 39:303–322.

Xue, X., Xie, X., Gunter, M., Rohan, T. E., Wassertheil-Smoller, S., Ho, G. Y. F., Cirillo, D., Yu, H., and Strickler, H. D. (2013). Testing the proportional hazards assumption in case-cohort analysis. *BMC Med. Res. Methodol.*, 13:88.

Yu, J., Tao, Y., Dou, J., Ye, J., Yu, Y., and Jin, L. (2018). The dose-response analysis between BMI and common chronic diseases in northeast china. *Sci. Rep.*, 8(1):4228.