Topics in Study Design and Analysis Involving Incomplete Data

by

Ce Yang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:

Dr. Lei Sun
Professor
Department of Statistical Sciences, Faculty of Art and Science
Division of Biostatistics, Dalla Lana School of Public Health
University of Toronto

Supervisor(s):

Dr. Liqun Diao
Assistant Professor
Department of Statistics and Actuarial Science
University of Waterloo

Dr. Richard J. Cook
Professor
Department of Statistics and Actuarial Science
University of Waterloo

Internal Members:

Dr. Mary Thompson
Distinguished Professor Emerita
Department of Statistics and Actuarial Science
University of Waterloo

Dr. Yeying Zhu
Associate Professor
Department of Statistics and Actuarial Science
University of Waterloo

Internal-External Member: Dr. Mark Oremus
Associate Professor
School of Public Health and Health Systems
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Incomplete data is a common occurrence in statistics with various types and mechanisms such that each can have a significant effect on statistical analysis and inference. This thesis tackles several statistical issues in study design and analysis involving incomplete data.

The first half of the thesis deals with the case of incomplete observations of the responses. In medical studies, events of interest are most likely to be under intermittent observation schemes, for example, detected via periodic clinical examinations. As a result, the event of interest is only known to happen within an interval, and the resulting interval-censored data hinders the application of numerous analysis tools. Although it is possible to presume the event time to happen at the endpoint or the midpoint of the interval, such *ad hoc* imputations are known to lead to invalid inferences. In Chapter 2, we propose appropriate imputations via censoring unbiased transformations and pseudo-observations of such incomplete responses to facilitate a straightforward use of prevalent machine learning algorithms. The former technique helps preserve the conditional mean structure with the presence of censoring, and the latter originates from the biased-corrected jackknife estimates. For a continuous response, both proposed imputations lead to regression trees models with the same expected L_2 loss as those fitted from complete observations. Therefore, prediction and variable selection naturally follow. Unlike most survival trees in literature, our proposed models do not rely on the widely made proportional hazard assumption. Furthermore, such models reduce to ordinary regression trees without the presence of censoring. Survivor function estimates of interval-censored data are required to employ the imputations; various semiparametric and nonparametric approaches are considered and compared. In particular, we scrutinize the case of current status data in a separate section.

The second half of the thesis addresses incomplete covariate data missing by design. Controlled by the investigators, the missingness is attributed to the budgetary constraints when measuring an "expensive exposure variable" in real-life scenarios. We focus on the well-known two-phase studies which exploit the response and inexpensive auxiliary information of the population to select a phase II sub-sample for the collection of the expensive

covariate. In Chapter 3, we look into an adaptive two-phase design that avoids the need for external pilot data. Dividing the phase II sub-sampling into multiple interim stages, we employ conventional sampling to select a fraction of the individuals of the phase II sub-sample to provide the information required for constructing an optimal sub-sample from those remaining to achieve maximum statistical efficiency subject to sampling constraints. Such adaptive two-phase designs naturally extend to multiple stages in phase II and are applicable when a surrogate of the exposure variable is available. Efficiency and robustness issues are investigated under various frameworks of analysis. As expected, the maximum likelihood approach that models the nuisance distribution tends to be more efficient, whereas inverse probability weighted estimating equations that avoid this tend to be more robust to the misspecification of the nuisance covariates models. The conditional maximum likelihood approach, to our delight, is well-balanced between the two. Moreover, the eagerness to gain efficiency while maintaining a certain level of robustness further drives us to explore semiparametric methods in all the analyses and designs.

Chapter 4 onward pays attention to more complicated settings in which covariates are missing in a sequence of two-phase studies with multiple responses and sampling constraints conducted on a common platform. For a given two-phase study, we expect to exploit not only information of the responses and auxiliary covariates at hand but also those passed on from earlier studies. We consider joint response models and perform secondary analyses of a new response using previously studied exposure variables. Moreover, the exposure variables acquired from earlier studies serve as pilot data to help construct an optimal selection model in an upcoming two-phase study. As we assess the balance between efficiency and robustness of the analysis methods, the potential misspecification of the joint response model warrants our attention. Finally, we note that the work can be extended to deal with two-phase response-dependent sampling with longitudinal data in Chapter 5.

Acknowledgements

I would like to express my gratitude to my supervisors, Professor Richard J. Cook and Professor Liqun Diao, for their kind support and excellent supervision. Their patience and guidance are indispensable for me to work on the thesis. I am exceptionally fortunate to have them as my supervisors.

I wish to thank the members of the examining committee, Professor Mary Thompson, Professor Yeying Zhu, Professor Mark Oremus, and Professor Lei Sun, for their insightful suggestions.

I wish to thank Prof. Grace Y. Yi for issuing me an offer of admission for the PhD program at the Department of Statistics and Actuarial Science, University of Waterloo.

I am grateful to Ker-Ai Lee for advice regarding statistical computing.

Special thanks go to Ms. Mary Lou Dufton for her kind support and communication during my graduate studies at Waterloo.

I should like to thank my peers: Xiyue Han, Chi-Kuang Yeh, Wenling Zhang, Xiaoxue Deng, Illia Sucholutsky, Qihuang Zhang, Haoxin Zhuang, Junhan Fang, Li-Pang Chen, Yechao Meng, Fangya Mao, Zhaohan Sun, Wenyuan Li, Trang Bui, Sheng Wang, Yiran Wang, Cong Jiang, Elinor Wang, and Xianwei Li. They are friendly, helpful, and contribute to a positive environment for learning.

I cherish my friend outside the department at Waterloo, too. I would like to thank Xi Dai and Minglei Li for being great roommates of mine.

Finally, I owe thanks to my family for encouraging me to pursue statistics at a graduate level. I have not been back home since my PhD program started in September 2018. It has been almost three years, and I do miss them.

Dedication

To Ms. Z. H. Qu.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis reports on the development and investigation of innovative statistical methodologies for dealing with incomplete data. There are two distinct themes to this research. The first is directed at the development of regression tree algorithms for interval-censored responses. The second is concerned with the development of two-phase designs for biomarker studies. Two projects within this second theme include approximate optimal design using adaptive sampling schemes and secondary analysis and sequential design of two-phase studies based on a common platform study. Sections 1.1 and 1.2 give brief reviews of the backgrounds of the themes. More detailed reviews can be found in relevant chapters.

1.1 Regression Trees for Interval-Censored Responses

Survival analysis is a branch of statistics involving methods for analyzing data where the response is the occurrence time of an event of interest, often referred to as a failure time. When the precise failure time is only known to lie in an interval, the resulting data are called interval-censored failure time data; See Sun (2006) for a thorough account of this field. Interval-censored data arise in many settings. For example, in a clinical trial, patients may visit the hospital periodically and the event of interest may occur between two consecutive visits. One interest is to develop predictive models based on a training

data set involving interval-censored data, which can be used to predict failure times and the failure status of individuals at a particular time horizon. A challenge is that the failure time in the test data set may also be subject to the interval censoring, and so the failure status may be unknown at the specified time horizon.

The predictive models we adapt to the interval censoring regime is the CART algorithm, one of the earliest and most popular methods for statistical learning that have seen considerable application in the past several decades (Loh, 2014). Splitting in a way that gives the greatest reduction in the sum of the within-node variances, the algorithm proposes a method of selecting an appropriate size of the tree. The graphical representation of the trees offers simple interpretation and is well-suited to the detection of effect modification typically addressed by incorporating interaction terms in regression models. Recent developments in CART include the extensions to handle a greater range of responses and methods for employing a more a variety of loss functions; see Loh (2014). For continuous responses, the CART algorithm returns regression trees which offer a convenient framework for the development of prediction algorithms wherein target values are calculated by the values in the terminal nodes of the regression trees. This appealing feature leads to scientific advances in many fields. For example, Henrichon and Fu (1969) refined the tree algorithm for application to problems in pattern recognition, while Meisel and Michalopoulos (1973) investigated the problem of space partitioning and piecewise constant approximations. The CART algorithm can be implemented by the `R` package `rpart` (Therneau and Atkinson, 2019).

While responses are interval-censored and therefore prohibit a straightforward application of the CART algorithm, the first theme of the thesis considers the censoring unbiased transformations and pseudo-observations to propose appropriate imputations of the responses. Under certain conditions, the CART algorithm following the imputations has a loss function equivalent to an unbiased estimator of the risk function when the data is complete. Such imputations, therefore, can be adapted to the regression trees instantly to handle interval-censored responses.

1.2 Optimal Two-Phase Designs

Two-phase designs can be traced back to the 1930s when Neyman (1938) tried to exploit the inexpensive auxiliary information from a population to estimate certain characteristics of interest. As the name suggests, two-phase designs naturally involve two stages. The first stage, denoted as phase I, collects the response and the auxiliary information for all individuals in the population at a low cost. The second phase, denoted as phase II, then selects a sub-sample to measure the characteristic of interest. It is expected that two-phase designs will be more efficient than simple random sub-sampling by exploiting the phase I data, especially when there is a strong association between the characteristic of interest and the auxiliary information. Introduced in the context of case-control studies by White (1982), two-phase designs have been widely used in epidemiological studies to control costs while providing efficient sampling strategies. The phase I sample is typically divided into strata according to the response and the inexpensive auxiliary information, followed by determining the sampling probabilities from each stratum in phase II for the measurement of some expensive information that we would not be able to afford to measure for the whole phase I sample. See Breslow and Cain (1988) and Scott and Wild (1991) for early work on fitting logistics regression models with two-phase data. Various analysis frameworks with different assumptions have been developed in the regime of two-phase designs including, but not limited to, maximum likelihood (Lawless et al., 1999; Breslow and Chatterjee, 1999), conditional maximum likelihood (Scott and Wild, 2011), mean score equations (Reilly and Pepe, 1995), inverse probability weighted estimating equations (Robins et al., 1994), and augmented inverse probability weighted estimating equations (Robins et al., 1994).

As far as a design is concerned, the higher the efficiency of the estimator of the parameter of interest, the better. However, the optimal design which yields maximum efficiency among those possible with the same expected phase II sub-sample size requires some *a priori* knowledge of the population model, including the parameters of interest. As a result, external pilot data that can be expensive or difficult to obtain is often required. This drawback was addressed by McIssac and Cook (2015) who advocated having multiple stages in phase II to avoid an external pilot study. Each stage uses information collected from the previous one and helps select the sampling strategy for the next. In practice, an adaptive

two-phase design can be implemented by dividing phase II into phase IIA and phase IIB. The former employs convenient sampling schemes to construct a phase IIA sub-sample for parameter estimation, and the latter uses the information obtained from phase IIA to construct an approximately optimal phase IIB sub-sample from the individuals who remain eligible. Both the sampling schemes employed in phase IIA and the ratio of the phase IIA sub-sample size to the phase IIB sub-sample size have been shown to affect the efficiencies of the adaptive two-phase designs.

The second theme of the thesis concentrates on the optimality of two-phase designs. Our work includes a thorough investigation of the properties of adaptive two-phase designs and scrutinizes their efficiency gains over non-adaptive ones under various analysis frameworks. In addition to the comprehensive exploration, the work extends beyond ordinary two-phase studies and considers the optimality of sequential two-phase designs involving multiple responses. Adopting joint response models, information of earlier two-phase studies helps achieve maximum statistical efficiency subject to budgetary constraints in the upcoming ones. The phase II sub-sample of the earlier study facilitates shaping the subsequent optimal design in the spirit of a phase IIA sub-sample in an adaptive two-phase design. Should there be no budget for an upcoming study, the joint response model enables us to utilize the exposure variables at hand to perform secondary analyses for a new response of interest.

1.3 Motivating Studies of Psoriatic Arthritis

Psoriatic arthritis (PsA) is a chronic, inflammatory joint disease associated with psoriasis, a long-lasting, noncontagious autoimmune disease characterized by raised areas of abnormal skin. While it can severely impact the quality of life, the disease progression of PsA is complex and heterogeneous. It is, therefore, desirable to identify patients at high risk of disease progression for early medical interventions. To obtain a better understanding of the illness, the Centre for Prognosis Studies in Rheumatic Disease at the Toronto Western Hospital launched in 1976 maintains a registry of patients with PsA, called the University of Toronto Psoriatic Arthritis Cohort (UTPAC), to study the disease progression (Chandran

et al., 2010b). The registry has been the largest cohort of patients with PsA in the world.

The projects of the thesis are inspired by various problems of the PsA research program, for which we give brief reviews as follows. More details in the application of the prediction models and two-phase designs to data from the UTPAC can be found in relevant chapters.

1.3.1 Prediction of Axial Disease

Chapter 2 of the thesis is motivated by the goal to predict axial disease, a chronic, inflammatory and degenerative condition of the axial skeleton (spine) that has a serious impact on functional ability and mobility among PsA patients. We consider a data set of a subset of the full registry of UTPAC comprised of 1022 patients undergoing biannual radiographic assessments. As a result, the failure times of the axial disease are interval-censored if the axial disease is found to present at one of the assessments. The data reflects the heterogeneity of timings of the assessments since the exact assessment times fluctuate across patients. Besides, some patients are found to miss some of the assessments. See Figure 1.1 for an illustration of the periodic assessments and the detection of axial disease. The data is heavily right-censored because about 62% of the patients did not develop the axial disease during the scheduled radiographic assessments. Information of their Human Leukocyte Antigen (HLA) biomarkers is available from a baseline biospecimen. Chapter 2 focuses on building regression tree models to identify HLA biomarkers useful for the prediction of axial disease within ten years of recruitment to the UTPAC. Previous research found the HLA biomarker *B27* significant to the development of the axial disease (Chandran et al., 2010b) and hence, serves as a reference for the evaluation of the performance of our proposed tree models.

1.3.2 An HLA Biomarker of Damage Progression

Chapter 3 is inspired by a separate study that looks into biomarkers associated with the development of bone damage over a short time horizon. As there is an increased interest in the design and analysis of biomarker studies related to PsA, patients in the registry

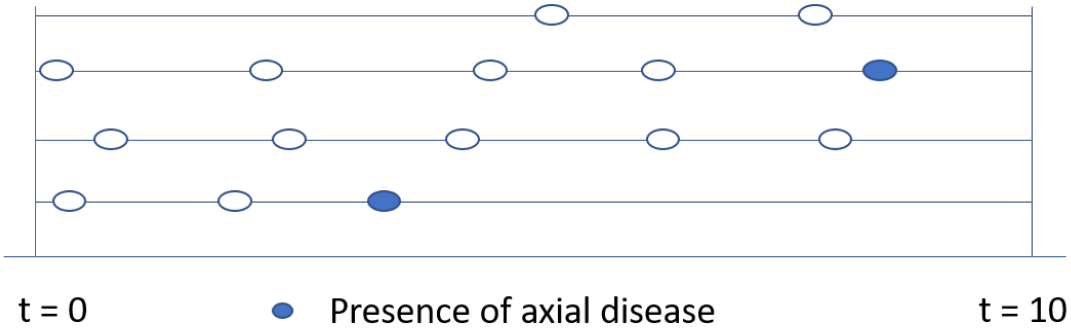

Figure 1.1: An illustration of periodic radiographic assessments (empty circle) for axial disease detection (solid circle) in ten years since recruitment.

have their blood samples drawn and stored in a bio-bank for examinations two years later to see if they have rapid disease progression. The interest lies in investigating the association between the biomarker matrix metalloproteinase 3 (MMP-3) and the progression of joint damage. It is cost-prohibitive to assay the blood samples for all patients in the registry. Still, levels of a traditional marker of inflammation, the erythrocyte sedimentation rate (ESR), are recorded at clinical visits (Gladman and Chandran, 2011) for all patients at a low cost. Chapter 3 focuses on developing adaptive two-phase designs to exploit the information from the ESR levels and hence best select a sub-sample of patients for measurements of MMP-3 from the cohort. See Figure 1.2 for an illustration of the problem.

We consider a data set of a subset of the full registry of UTPAC comprised of 251 patients with MMP-3 measurements available at a baseline assessment. While the MMP-3 sample date is chosen as the baseline assessment date, the number of the damaged joints is defined as the number obtained from the medical assessment closest to the MMP-3 sample date. The disease progression status is determined by the increment of the number of damaged joints within two years. The data is used as pilot data to help inform parameter configurations for a focused simulation study to evaluate the performance of adaptive two-phase designs in the setting of the PsA study.

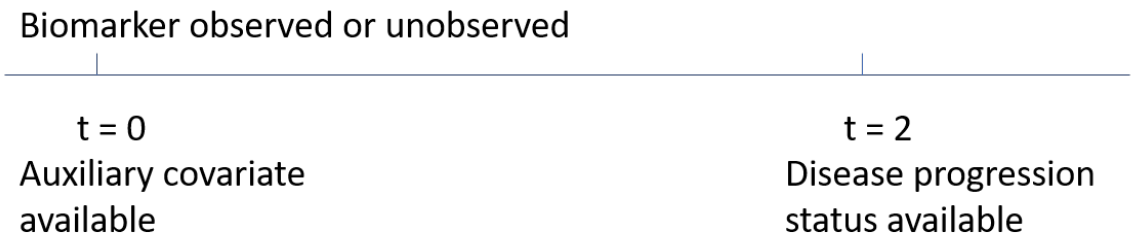

Figure 1.2: An illustration of the information of disease progression status, biomarker, and auxiliary covariate in two years of follow-up.

1.3.3 Secondary Use of HLA Biomarker Data

Chapter 4 addresses the need of the researchers to use a previously studied biomarker to enhance a subsequent two-phase study with a new response. We consider a data set of a subset of the full registry of UTPAC comprised of 706 patients. Other than the auxiliary information, some patients have their biospecimens assayed to inspect the relationship between an HLA biomarker and the progression of clinical joint damage. It is favourable to utilize these available biomarkers to analyze new responses, such as the development of active joints (swollen joints or joints losing range of motion with pain or tenderness) to save costs and preserve biospecimens. Chapter 4 develops methods for secondary analyses of the progression of active joints without additional assays in the context of two-phase studies. Moreover, we propose sequential two-phase designs to exploit information passed from the clinical joint damage study to inform the optimal sampling strategy of the upcoming active joints study subject to new budgetary constraints. A valid aggregation of the biomarkers measured from the two sequential studies is appealing as a larger combined phase II subsample leads to more efficiency. See Figure 1.3 for an illustration of two such sequential two-phase studies (Study 1 and Study 2) conducted on the same platform.

1.4 Outline of Thesis

The remainder of the thesis is organized as follows.

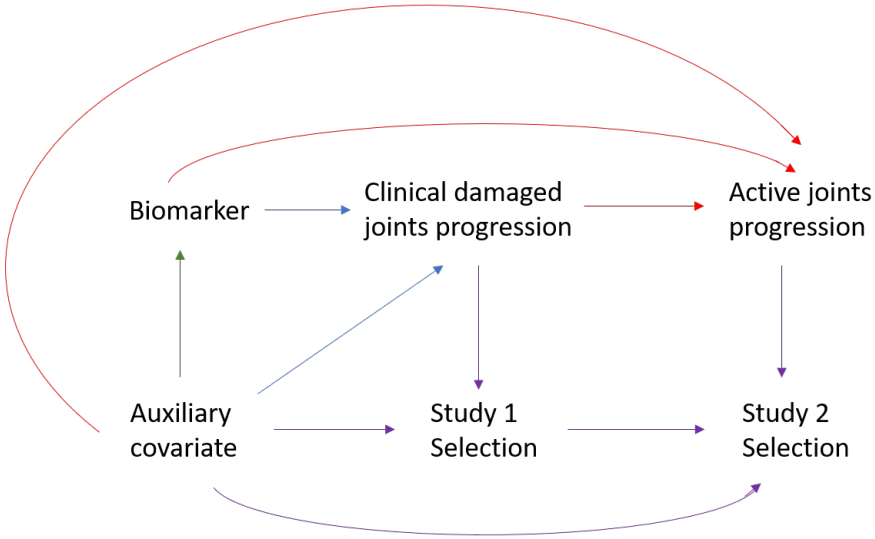

Figure 1.3: An illustration of the information of two sequential two-phase studies (Study 1 and Study 2) with the same biomarker and different responses.

Chapter 2 discusses building regression trees for interval-censored responses by carefully constructing appropriate observed data loss functions. We illustrate the equivalence between such observed data loss functions and imputed loss functions based on censoring unbiased transformations and pseudo-observations. The proposed regression trees are advantageous in predicting failure times and failure status in test data sets and are good at recovering underlying tree structures. Survival trees for current status data, a special case of interval-censored data, are available, too. An application is given to a study involving PsA patients where the aim is to determine the influential predictors and predict the status of the axial disease at a specific time in future.

Chapter 3 focuses on the efficiency and robustness of adaptive two-phase designs. We investigate different methods of analysis in the framework of adaptive two-phase designs and present sampling strategies that are both asymptotically efficient and robust to model misspecification. Semiparametric analysis frameworks are found to play a role in the design stage to enhance the robustness without suffering obvious efficiency loss when the exposure variable is continuous. The designs can be adapted to the surrogate value problem. An application is given to a study involving PsA patients where the aim is to determine how

to best select individuals for the expensive measurements of the genetic markers associated with the development of the disease.

Chapter 4 turns focus to large cohort studies supporting a series of biomarker studies addressing distinct but sometimes related scientific questions. We extend the perspective of secondary analysis in case-control studies to deal with a sequence of two-phase designs. Leveraging the available exposure data acquired from earlier studies, we examine the relationship between a previously studied biomarker and a new response. We also consider the design of two-phase studies that exploit the information available from previous two-phase studies conducted on the same platform. Using joint response models, we consider and compare likelihood methods with inverse probability weighted estimating equations in efficiency and robustness. An application is given to a study involving PsA patients where the aim is to make secondary use of previously assayed biospecimens to study the progression of related diseases.

Finally, Chapter 5 reviews the contributions of the thesis and outlines future research topics, including a layout of two-phase designs with longitudinal responses.

Chapter 2

Regression Trees for Interval-Censored Responses Based on Censoring Unbiased Transformations and Pseudo-Observations

2.1 Introduction

Regression trees have been used extensively for prediction problems involving failure time data where new splitting criteria and evaluation metrics have been required to deal with right-censored observations. Gordon and Olshen (1985) first attempted to obtain Kaplan-Meier estimates (Kaplan and Meier, 1958) to the data in the nodes and use the distance measures between the within-node fitted curves as the splitting criterion. Davis and Anderson (1989) considered exponential log-likelihood as the loss function and proposed an "exponential tree" to analyze the covariates effects for right-censored data. The "exponential tree" quantifies the prediction error via the true and estimated hazard functions.

Adopting most aspects of the CART algorithm, LeBlanc and Crowley (1992) developed a method to obtain the estimates of the relative risks for right-censored survival data. This generalization of the proportional hazard regression to relative risk functions is done by fitting a proportional hazard model in the nodes and maximizing the reduction in one-step deviance during the splitting. LeBlanc and Crowley (1993) proposed the idea of using the dissimilarity in the survival distributions of patients as the splitting criterion which is measured by the log-rank test. Molinaro et al. (2004) proposed a method to handle right censoring by replacing the complete data loss function with an unbiased observed data loss function and used inverse probabilities of censoring weights when constructing the regression trees.

In many settings interest lies in an event that is not directly observable but can be detected to have occurred upon careful clinical examination or through the use of laboratory tests or imaging. Examples include the development of asymptomatic vertebral fractures in patients with osteoporosis (Cano et al., 2016), the development of new lesions in skin cancer studies (Abu-Libdeh et al., 1990), and the development of skeletal metastases in breast cancer (Hortobagyi et al., 1996). When event times of interest are under intermittent observation schemes, methods for handling interval-censored failure time data are required. Should there be only a single assessment time, it is favourable to develop statistical methods specifically for current status data. Sun (2006) gave a thorough account of this field, describing methods for parametric analyses and various types of semiparametric analyses.

Frameworks for predictive modelling are not as well-developed for interval-censored data as they are for right-censored data. Yin and Anderson (2002) proposed a regression tree algorithm based on log-likelihood for interval-censored data when assuming that the failure time follows an exponential distribution, which extends the "exponential tree" proposed by Davis and Anderson (1989) for right-censored data. In Yin and Anderson (2002), the authors further proposed a nonparametric tree for interval-censored data, a regression tree algorithm based on the nonparametric maximum likelihood estimator (NPMLE) in which probabilities that an event occurs in the innermost intervals (Yu et al., 2000) are estimated using the self-consistent algorithm (Turnbull, 1976). Yin and Anderson (2002) commented that the nonparametric tree performs reasonably well regardless of the true underlying failure time distribution, especially when the sample size is large and the dropout rate is low.

However, the exponential tree does not perform well when there is an appreciable trend in the hazard. Fu and Simonoff (2017) proposed another nonparametric tree algorithm under the framework of the conditional inference tree (Hothorn et al., 2006), which addressed the problem of variable selection bias by separating the selection of the splitting variables and the splitting points into two steps. The algorithm is based on the log-rank score for interval-censored data with the survivor functions estimated by Turnbull's self-consistent algorithm. Empirically, the conditional inference tree method of Fu and Simonoff (2017) showed better predictive performance for interval-censored data compared to alternative *ad hoc* approaches such as imputing the event times using the left endpoint, mid-point, or right endpoint of the censoring interval. Wu and Cook (2020) discussed and evaluated methods for fitting and assessing the predictive accuracy of models when training and validation data feature interval-censored failure times. As for current status data, there is no predictive model approach developed specifically for this type of interval-censored data to our knowledge.

In this chapter, we propose algorithms for developing regression trees for interval-censored failure times by constructing observed data loss functions which are consistent estimators of a complete data risk function. We demonstrate that regression trees built using the L_2 observed-data loss functions are equivalent to the ones obtained by applying the complete data regression trees to the imputed failure times. We discuss strategies to construct observed data loss functions and target quantities for prediction including a failure time and whether a subject has experienced failure by a particular time. The proposed methods are evaluated empirically and compared to the oracle tree built using uncensored failure times, methods based on *ad hoc* imputation approaches, and the conditional inference tree approach proposed by Fu and Simonoff (2017).

2.1.1 A Review of the Classification and Regression Trees Algorithm

Before we start, we give a brief review of the Classification and Regression Trees (CART) algorithm (Breiman et al., 1984). A decision tree is a hierarchically organized structure of nodes and branches. Figure 2.1 provides an illustrative example of a Classification and

Regression Tree. As a basic unit of a tree structure, a node contains a subset of the data used to construct the learning algorithm. The root node (node 1) is the node on the top of the tree, formed by the entire data set. The two branches underneath node 1 indicate the split of node 1 and the splitting leads to two children nodes (nodes 2 and 3). The parent of a node is its immediate predecessor node and the children of a node are its immediate successors. The nodes which do not have children are the terminal nodes of the tree (nodes 3, 4 and 5). The terminal nodes form a partition of the data set.

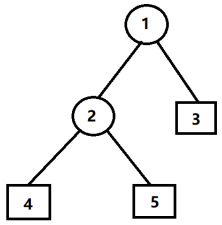

Figure 2.1: An illustration of a classification and regression tree.

Let $(Z, X')'$ denote a data vector, where Z represents a scalar-valued response with support on a subset of the real line $\mathbb{R}$, and $X \in \mathcal{X} \subset \mathbb{R}^p$ represents a p-dimensional vector of covariates. Let $\mathcal{D} = \{Z_i, X_i : i = 1, \ldots, n\}$ be the data set containing n independent and identically distributed (i.i.d.) copies of the data. Define a prediction rule as $\Psi(X) : \mathcal{X} \to \mathbb{R}$, a real-valued function. If the prediction function $\Psi(X)$ takes a piece-wise constant form, the prediction rule can be written as

$$\Psi(X) = \sum_{k=1}^{K} \beta_k I(X \in \mathcal{X}_k), \tag{2.1}$$

for any $X \in \mathcal{X}$, where $\{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ is a finite partition of the covariates space $\mathcal{X}$, i.e., $\mathcal{X}_k \cap \mathcal{X}_j = \phi$ for $k \neq j$, $k, j = 1, \ldots, K$, and $\cup_{k=1}^{K} \mathcal{X}_k = \mathcal{X}$; β_k is the predicted value if X falls into the kth partition $\mathcal{X}_k$, for $k = 1, \ldots, K$. The algorithm aims to train a prediction function $\Psi(X)$.

To quantify the quality of the prediction rule, a loss function $L(Z, \Psi(X))$ is defined as a nonnegative measure of the distance between the response Z and its prediction $\Psi(X)$. The risk is defined as the expected loss, i.e., the expectation of the loss function $\mathcal{R}(\Psi) = E[L(Z, \Psi(X))]$. The optimal prediction function is the one which minimizes the risk. A common choice of the loss function is the squared error loss (or L_2 loss) function $L_2(Z, \Psi(X)) = (Z - \Psi(X))^2$. When an L_2 loss function is used, the optimal prediction function is the conditional mean $\Psi_0(X) = E(Z|X)$ obtained by minimizing the L_2 risk $\mathcal{R}(\Psi) = E[L_2(Z, \Psi(X))]$. The average loss of a subset $\mathcal{A}$ of the data set $\mathcal{F}$ ($\mathcal{A} \in \mathcal{F}$) is defined as

$$L(\mathcal{A}, \Psi) = \frac{1}{\sum_{i=1}^{n} I((Z_i, X_i')' \in \mathcal{A})} \sum_{i=1}^{n} I((Z_i, X_i')' \in \mathcal{A}) L(Z_i, \Psi(X_i)).$$

The CART algorithm, like many other machine learning algorithms, is designed to search for an optimal prediction rule.

A CART algorithm proceeds by growing a large tree, applying a cost-complexity pruning, and using a cross-validation procedure to select the "optimal"-sized tree. The detailed procedure is summarized as follows. First, the algorithm builds a large tree by iterative binary splitting. The algorithm starts with the root node. To find the best split at a current node, we identify all possible binary splits and choose the combination of a splitting covariate and a splitting point such that there will be the greatest reduction in the total loss of the current node (i.e., the sum of the loss of all children of the current node). Then the algorithm splits the current node into two children nodes on the selected splitting covariate at the selected splitting point. Iterating this procedure, a tree structure is formed by iteratively splitting nodes to maximize the decrease in the total loss until pre-determined stopping criteria are met. The stopping criteria are usually based on constraints on the structure of the tree. The depth of a tree is defined as the maximal levels of nodes from the root node to a terminal node. The `minsplit` is defined as the minimum number of observations that must exist in a node for a split to be attempted. The `minibucket` is defined as the minimum number of observations in any terminal node. Once the stopping criteria are met, the algorithm stops splitting further and returns a large tree, denoted by Φ_{max}. Popular stopping criteria set a maximum tree at a depth of 30, `minsplit` of 20 and `minibucket` of 7 (the default setting for `rpart` package in `R`).

The second step involves pruning the tree by the introduction of a cost-complexity penalty function. For a data set of n independent individuals, $i = 1, \ldots, n$, the cost-complexity function is defined as

$$K_\alpha(\Phi) = \sum_{i=1}^{n} L(Z_i, \Psi_\Phi(X_i)) + \alpha|\Phi|,$$

where Φ represents the tree structure, Ψ_Φ is the predictor obtained from tree Φ, α is a non-negative real tuning parameter penalizing the cost via the size of the tree, and $|\Phi|$ is the number of the terminal nodes in the tree. For each α, the idea is to find the subtree Φ_{max}, which minimizes the cost-complexity function. Although α runs through a continuum of values, there are at most a finite number of sub-trees of Φ_{max}. To see this, note that if $\Phi^{(\alpha)}$ denotes the subtree that minimizes K_α for a given α, then it keeps minimizing K_α as α increases until a jump point α' is reached, where a new subtree $\Phi^{(\alpha')}$ becomes the tree with minimal K_α until the next jump point α'', and so on. In other words, starting with $\alpha_1 = 0$ and $\Phi^{(\alpha_1)}$, the smallest sub-tree of Φ_{max} that minimizes K_0, we can find a decreasing sequence of sub-trees $\Phi^{(\alpha_k)}$ for $k \geq 1$, which correspond to an increasing sequence of α_k such that $\Phi^{(\alpha_k)}$ is the smallest sub-tree minimizing the cost-complexity function K_α for $\alpha_k \leq \alpha < \alpha_{k+1}$. The obtained sequence of sub-trees, denoted by $\Phi^{(\alpha_1)}, \ldots, \Phi^{(\alpha_L)}$, are the candidates for the "optimal" tree, where L is the number of sub-trees of Φ_{max} in the sequence.

The CART algorithm uses cross-validation to choose the "optimal" tree. In a V-fold cross-validation, the data are randomly split into V mutually exclusive folds with as near as possible equal size; common choices of V are 5 or 10, corresponding to 5-fold and 10-fold cross-validation. We repeat the tree growing and pruning procedure using the data excluding fold v, $v = 1, \ldots, V$. Thereby, for each $v = 1, \ldots, V$, a decreasing sequence of sub-trees $\Phi_v^{(\alpha_l)}$, for $l = 1, \ldots, L$, is obtained. Let $S_{i,v}$ indicate whether subject i belongs to fold v. We then define the cross-validated estimator of risk as

$$RCV(\alpha_l) = \frac{1}{n} \sum_{v=1}^{V} \sum_{i=1}^{n} I(S_{i,v} = 1) L(Z_i, \Psi_v^{(\alpha_l)}(X_i)),$$

where $\Psi_v^{(\alpha_l)}(X)$ is the prediction function from tree $\Phi_v^{(\alpha_l)}$. For each $l = 1, \ldots, L$, calculate $RCV(\alpha_l)$ and let l_{max} denote the value of l which minimizes $RCV(\alpha_l)$. The "optimal"

tree is $\Phi^{(\alpha_{l_{max}})}$, the sub-tree of the maximum tree Φ_{max} obtained from the tree growing procedure. The prediction rule obtained by the CART algorithm is the piecewise function in (2.1) according to the partition formed by the terminal nodes of the "optimal" tree $\Phi^{(\alpha_{l_{max}})}$.

2.1.2 A Review of the Conditional Inference Trees

Here follows a brief review of the conditional inference tree framework (Fu and Simonoff, 2017; Hothorn et al., 2006) as it will be the main competitor of our proposed regression trees in the simulation studies. The framework focuses on the conditional distribution of the response given the covariates in the context of tree-structured recursive partitioning. Different from the CART algorithm, the conditional inference trees first determine the splitting variables and then decide the specific splitting points.

Given a learning sample $L_n = \{Z_i, X_{i1}, \ldots, X_{ip} : i = 1, \ldots, n\}$, the recursive partitioning is formulated with the help of integer-valued case weights $\nu = (\nu_1, \ldots, \nu_n)'$. Each node of a conditional inference tree can be represented by the case weights which are non-zero if and only if the corresponding data item belongs to the node. Hothorn et al. (2006) assumed the weights to be either zero or one for convenience. For each node identified by the case weights, the splitting decision is made by information of the response Z covered by the covariates X. This is achieved by considering the partial hypothesis on the conditional distribution

$$H_0^j : P(Z|X_j) = P(Z), j = 1, \ldots, p$$

followed by the global independence null hypothesis

$$H_0 = \bigcap_{j=1}^{p} H_0^j.$$

If H_0 cannot be rejected at a prespecified level (0.05 by default), the recursion stops. Otherwise, the splitting variable is determined by the association between Z and each covariate X_j tested in the partial hypothesis H_0^j, $j = 1, \ldots, p$ with linear statistics of the form

$$T_j(L_n, \nu) = \text{vec}\left[\sum_{i=1}^{n} \nu_i g_j(X_{ij}) h(Z_i)'\right],$$

where g_j is a nonrandom transformation of covariate X_j, h is the influence function of the response, and the resulting matrix is combined to a column vector by the *vec* operation. See Hothorn et al. (2006) for more details in choices of g and h under various circumstances. Finally, the linear statistics is standardized using the conditional distribution and covariance under H_0 computed in Strasser and Weber (1999). Under the null hypothesis, asymptotic normality follows. The covariate associated with the smallest p-value when compared to the corresponding normal quantiles is chosen as the splitting variable.

For a given splitting variable, the splitting point can be established by any splitting criteria, including those for the CART algorithm reviewed in Section 2.1.1. In Hothorn et al. (2006), the goodness of a split is assessed by the two-sample statistic induced by the linear statistic

$$T_j^{\mathcal{A}}(L_n, \nu) = \text{vec}\left[\sum_{i=1}^{n} \nu_i I(X_{ij} \in \mathcal{A}_j) h(Z_i)'\right],$$

where $\mathcal{A}_j$ is a subset of the sample space $\mathcal{X}_j$, and the best split is the one that maximizes the resulting standardized test statistic. The conditional inference tree is formed by repeating steps in choosing splitting variables and splitting points until the global independence null hypothesis cannot be rejected. The separation of choosing splitting variables and splitting points avoid a systematic tendency towards the covariates with many possible splits.

The remainder of the chapter is organized as follows. In Section 2.2, we develop a strategy for the construction of observed data loss functions with censored responses, followed by implementing the CART algorithm based on these functions. In Section 2.3, we assess the performance of our methods empirically by how well they recover the correct tree structure and based on predictive performance. We also compare the proposed methods with methods based on *ad hoc* imputation strategies and the conditional inference tree approach of Fu and Simonoff (2017). In Section 2.4 we report on an analysis aiming to identify biomarkers associated with the onset of axial disease in patients with psoriatic arthritis; we also predict axial involvement ten years from the baseline assessment. In Section 2.5, we adapt our strategies to develop survival trees for current status data. Concluding remarks and topics of future research are given in Section 2.6.

2.2 CART for Interval-Censored Failure Time Data

Here we provide a general framework for handling censored data which facilitates a generalization of the CART algorithm to accommodate interval-censored responses. We introduce two strategies to construct observed data loss functions and use them to replace the complete data loss function for uncensored data.

2.2.1 Notation and Preliminaries

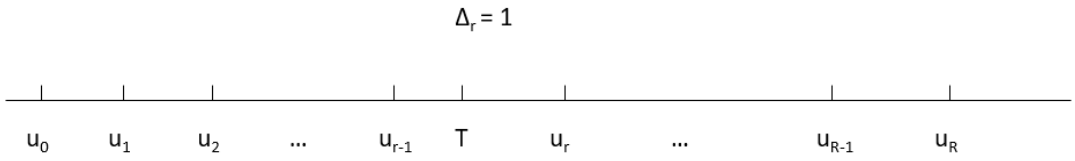

Figure 2.2: An illustration of interval-censored failure time data.

From here till the end of the chapter, let $T \in \mathbb{R}^+$ denote the failure time of interest and $S(t|x) = P(T > t|X = x)$ denote the conditional survivor function of T given covariates $X = x$. Furthermore, let $Z = h(T)$, a transformation of the failure time, where $h(\cdot)$ is a known monotone increasing function. Examples of $h(\cdot)$ include the identity function or $h(u) = I(T > u)$, an indicator of whether T is greater than some time point u of interest.

Interval censoring is a form of coarsening wherein the failure time T is only known to lie in an interval. Type R interval censoring arises if R assessment times $u_1 < u_2 < \cdots < u_R < \infty$ are available beyond $u_0 = 0$, along with functions $\Delta_r = I(u_{r-1} < T \leq u_r)$ indicating whether the failure time T lies in the rth finite interval $(u_{r-1}, u_r]$, $r = 1, \ldots, R$; we let $\Delta_{R+1} = 1 - \Delta_1 - \cdots - \Delta_R = I(T > u_R)$. The observed data of type R interval censoring are thus $O = (u_1, u_2, \ldots, u_R, \Delta_1, \Delta_2, \ldots, \Delta_R)$. When $R = 1$, there is only one inspection time $U > 0$ and the special case of interval censoring gives rise to current status data; we consider such data in Section 2.5. We further assume that conditional on the covariates that are being controlled for, the failure times are mutually independent of the assessment process as in Cook and Lawless (2019); when the $u_1, \ldots, u_R$ are fixed and prescheduled this

is of course the case. With a sample of data on n independent processes we label individuals with the subscript i and denote the observed data as $\mathcal{O} = \{O_i, X_i : i = 1, \ldots, n\}$, which contains information on realizations of the n i.i.d. joint processes.

2.2.2 Construction of the Observed Data Loss Functions

Recall that $\Psi(X) : \mathcal{X} \to \mathbb{R}$ denotes the prediction rule. The loss function one would use with complete data $\mathcal{D} = \{Z_i, X_i : i = 1, \ldots, n\}$ is $L(\mathcal{D}, \Psi) = \frac{1}{n}\sum_{i=1}^{n} L(Z_i, \Psi(X_i))$ and $\mathcal{R}(\Psi) = E[L(Z, \Psi(X))]$ is the complete data risk. Our goal is to build regression trees based on interval-censored data, in which case the Z_i's are not observable and the complete data loss function $L(\mathcal{D}, \Psi)$ cannot be calculated. To address this we define a class of *observed data loss functions* $L(\mathcal{O}, \Psi)$ to be used in place of $L(\mathcal{D}, \Psi)$ in the tree growing, pruning and cross-validation steps in CART algorithm. We choose such functions so that the observed data loss function is an unbiased or consistent estimator of the complete data risk $\mathcal{R}(\Psi)$.

We next propose two strategies of constructing observed data loss functions, including censoring unbiased transformations and methods based on pseudo-observations.

Observed Data Loss via Censoring Unbiased Transformations

Review of CUT and Extensions to Accommodate Interval Censoring

Censoring unbiased transformations (CUT) have been utilized to deal with right-censored data; see Fan and Gijbels (1996) and Rubin and van der Laan (2007). The Buckley-James transformation is an example of a CUT that was proposed to facilitate linear regression for right-censored data (Buckley and James, 1979). More recently, Steingrimsson et al. (2019) consider the constructions of observed data loss functions using CUTs for building regression trees with right-censored data. Here we generalize the definition of CUT to accommodate interval-censored data. Let $\mathscr{Y}$ be a scalar function of the complete data $(Z, X')'$ and $\mathscr{Y}^*$ be a scalar function of the observed data $(O, X')'$. We define $\mathscr{Y}^*$ as a CUT for $\mathscr{Y}$ if

$$E[\mathscr{Y}^*(O, X)|X = x] = E[\mathscr{Y}(Z, X)|X = x]$$

for every $x \in \mathcal{X}$.

General Construction of the Observed Data Loss via CUT

By setting $\mathscr{Y}(Z, X) = L(Z, \Psi(X))$, $\mathscr{Y}^*(O, X)$ is a CUT of $L(Z, \Psi(X))$. For a sample of n independent individuals the observed data loss function $L(\mathcal{O}, \Psi)$ is constructed using the empirical average of the $\mathscr{Y}^*(O, X)$ terms as

$$L_{CUT}(\mathcal{O}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{Y}^*(O_i, X_i) \ . \tag{2.2}$$

The constructed observed data loss function (2.2) is thus an unbiased estimator for the complete data risk $\mathcal{R}(\Psi) = E[L(Z, \Psi(X))]$.

Constructing the L_2 Observed Data Loss via CUT

When building a CART with complete data, the default complete data loss for a continuous response is the L_2 loss and a piecewise constant prediction rule $\Psi(X) = \sum_{k=1}^{K} \beta_k I(X \in \mathcal{X}_k)$ is adopted, where $\{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ is a finite partition of the covariate space $\mathcal{X}$ and β_k is the predicted value if X falls into the kth partition $\mathcal{X}_k$, for $k = 1, \ldots, K$. When the complete data loss takes the form

$$L_2(Z, \Psi(X)) = \sum_{k=1}^{K} I(X \in \mathcal{X}_k)(Z^2 - 2Z\beta_k + \beta_k^2) \ , \tag{2.3}$$

the observed data loss function is built using the empirical average of its corresponding CUTs given by

$$L_{2,CUT}(\mathcal{O}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(X_i \in \mathcal{X}_k) \left[\mathscr{Y}_2^*(O_i, X_i) - 2\mathscr{Y}_1^*(O_i, X_i)\beta_k + \beta_k^2\right] \ , \tag{2.4}$$

where $\mathscr{Y}_1^*(O, X)$ and $\mathscr{Y}_2^*(O, X)$ are the CUTs for Z and Z^2, respectively. The expression in (2.4) clearly has the same conditional expectation as the L_2 complete data loss (2.3) given covariates X, so (2.4) is an unbiased estimator of the complete data risk $\mathcal{R}(\Psi) = E[L_2(Z, \Psi(X))]$.

Censoring Unbiased Transformations (CUTs) for Z^j

The challenge reduces to finding a suitable function $\mathscr{Y}_j^*(O, X)$, the CUT for Z^j, $j = 1, 2$. Since

$$E(Z^j|X) = \sum_{r=1}^{R+1} E(Z^j|\Delta_r = 1, X)P(\Delta_r = 1|X),$$

the CUT of Z^j can be constructed as

$$\mathscr{Y}_j^*(O, X) = \sum_{r=1}^{R+1} \Delta_r E(Z^j|\Delta_r = 1, X), \tag{2.5}$$

where $Z = h(T)$ and

$$E(Z^j|\Delta_r = 1, X) = \frac{1}{S(u_r|X) - S(u_{r-1}|X)} \int_{u_{r-1}}^{u_r} h(t)^j dS(t|X),$$

for $r = 1, \ldots, R + 1$. In this case $\mathscr{Y}_j^*(O, X)$ has the same conditional expectation as Z^j given covariates X (i.e., it is a CUT of Z^j). When interest lies in the failure time itself (i.e., $Z = T$),

$$E(Z^j|\Delta_r = 1, X) = \frac{u_r^j S(u_r|X) - u_{r-1}^j S(u_{r-1}|X) - \int_{u_{r-1}}^{u_r} S(t|X) j t^{j-1} dt}{S(u_r|X) - S(u_{r-1}|X)}, \tag{2.6}$$

while when interest lies in the failure status at a fixed time u (i.e., $Z = I(T > u)$),

$$\begin{aligned} E(Z^j|\Delta_r = 1, X) &= \frac{1}{S(u_r|X) - S(u_{r-1}|X)} \{I(u \le u_{r-1})[S(u_r|X) - S(u_{r-1}|X)] \\ &\quad + I(u_{r-1} < u \le u_r)[S(u_r|X) - S(u|X)]\} \\ &= I(u \le u_{r-1}) + I(u_{r-1} < u \le u_r) \frac{S(u_r|X) - S(u|X)}{S(u_r|X) - S(u_{r-1}|X)}. \end{aligned} \tag{2.7}$$

The conditional survivor function of T given covariates X can be estimated semiparametrically under a Cox proportional hazard model or nonparametrically using the conditional inference trees proposed Fu and Simonoff (2017). In addition, we hereby point out the possibility of using the marginal survivor function of T directly to construct a counterpart of the CUT. Turnbull (1976) helps with the estimation; see 2.2.3 for details.

Building the Observed Data Loss via Pseudo-Observations

Review of the Pseudo-Observation Approach

The jackknife pseudo-observation (PO) approach for incomplete data was introduced and originally used in standard regression settings (Quenouille, 1949; Tukey, 1958), but has been greatly promoted for applications to survival analysis in recent years; see Andersen and Perme (2010) for a recent review. Andersen et al. (2003) applied the pseudo-observation approach for inferences based on multi-state models; Andersen et al. (2004) used the pseudo-observations in a regression of restricted mean survival time with right-censored data; and Han et al. (2014) used the pseudo-observations to a semiparametric regression for interval-censored responses.

Suppose $\hat{\theta}$ is an estimator of a parameter of interest θ based on an i.i.d. sample of $Z_1, \ldots, Z_n$, and $\hat{\theta}^{(-i)}$ is a leave-one-out estimator of θ based on $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$. The ith pseudo-observation is constructed as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)}, \tag{2.8}$$

for $i = 1, \ldots, n$. If $\hat{\theta}$ and $\hat{\theta}^{(-i)}$ are unbiased estimators of θ, the expectation of $\hat{\theta}_i$ is equal to θ and thus the empirical average of POs also gives an unbiased estimator of θ.

Constructing the Observed Data Loss via Pseudo-Observations (POs)

We aim to construct an observed data loss function which is unbiased for the full data risk $\mathcal{R}(\Psi)$. Thus, we set the quantity of interest $\theta = \mathcal{R}(\Psi)$. Suppose that $\hat{\theta}$ is an estimator of θ using the observed interval-censored data $\mathcal{O}$ and $\hat{\theta}^{(-i)}$ is the corresponding leave-one-out estimator using $\mathcal{O}^{(-i)} = \{O_j, X_j : j = 1, \ldots, i-1, i+1, \ldots, n\}$. The POs are obtained using (2.8), and they are used to further construct the observed data loss function

$$L_{PO}(\mathcal{O}, \Psi) = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i.$$

Constructing the L_2 Observed Data Loss via Pseudo-Observations

The special case when the L_2 loss and piecewise constant prediction rules are adopted warrants special consideration. The observed data loss function built using the empirical average of the POs is

$$L_{2,PO}(\mathcal{O}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(X_i \in \mathcal{X}_k) \left(\hat{\theta}_{2i} - 2\hat{\theta}_{1i}\beta_k + \beta_k^2 \right), \tag{2.9}$$

where $\hat{\theta}_{2i}$ and $\hat{\theta}_{1i}$ are the POs for Z_i^2 and Z_i in the complete data loss (2.3). If $\hat{\theta}_j$ is an estimator of $E(Z^j)$ and $\hat{\theta}_j^{(-i)}$ be the corresponding leave-one-out estimator, then $\hat{\theta}_{ji} = n\hat{\theta}_j - (n-1)\hat{\theta}_j^{(-i)}$ is the ith PO for $E(Z^j)$, $j = 1, 2$.

The Pseudo-Observations for $E(Z^j)$

Since $Z = h(T)$, we estimate $E(Z^j) = -\int_0^\infty h(t)^j dS(t)$ by replacing $S(t)$ with an estimate so that the POs for $E(Z^j)$ can be written as

$$\hat{\theta}_{ji} = -n \int_0^\infty h(t)^j d\hat{S}(t) + (n-1) \int_0^\infty h(t)^j d\hat{S}^{(-i)}(t) \,,$$

where $\hat{S}(\cdot)$ is an estimator of the survivor function $S(\cdot)$, and $\hat{S}^{(-i)}(\cdot)$ is the corresponding leave-one-out estimator excluding data from individual i.

When interest lies in the failure time,

$$\begin{aligned} \hat{\theta}_{1i} &= n \int_0^\infty \hat{S}(t)dt - (n-1) \int_0^\infty \hat{S}^{(-i)}(t)dt, \\ \hat{\theta}_{2i} &= 2 \left[n \int_0^\infty t\hat{S}(t)dt - (n-1) \int_0^\infty t\hat{S}^{(-i)}(t)dt \right] \,, \end{aligned} \tag{2.10}$$

whereas when interest lies in the failure status at a fixed time u we have

$$\hat{\theta}_{ji} = n\hat{S}(u) - (n-1)\hat{S}^{(-i)}(u), \tag{2.11}$$

for $j = 1, 2$. The Turnbull algorithm (Turnbull, 1976) can be used to obtain a nonparametric estimator of the marginal survivor function when the failure times are subject to interval censoring. We defer the discussion on nonparametric maximum likelihood estimation for interval-censored data to the end of this section.

Response Imputation

Review of Response Imputation

The idea of response imputation was introduced in Steingrimsson et al. (2019) to facilitate a straightforward use of the observed data loss function for the CART algorithm when data are right-censored. Theorem 4.1 in Steingrimsson et al. (2019) implied that one can implement the L_2 observed data loss functions by applying the L_2 complete data CART algorithm with some imputed data set $\mathcal{Z} = \{\hat{Z}(O_i, X_i), X_i : i = 1, \ldots, n\}$.

Note that using the complete data CART algorithm following imputation is equivalent to using the imputed loss function. To see this, the imputed loss function is

$$L_{2,I}(\mathcal{Z}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(X_i \in \mathcal{X}_k) \left[\hat{Z}(O_i, X_i)^2 - 2\hat{Z}(O_i, X_i)\beta_k + \beta_k^2 \right], \tag{2.12}$$

where $\hat{Z}(O_i, X_i)$ is the imputed response for the ith subject in the data set. Theorem 4.1 in Steingrimsson et al. (2019) showed that the CART algorithm makes decisions in the tree growing, pruning and cross-validation steps, which do not depend on the term $\hat{Z}(O_i, X_i)^2$.

Response Imputation for Interval-censored Data

We extend the idea of response imputation for interval-censored data. We can make the imputed loss function (2.12) equivalent to the L_2 CUT loss function (2.4) by letting $\hat{Z}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$; or implementing the L_2 PO loss function (2.9) by letting $\hat{Z}(O_i, X_i) = \hat{\theta}_{1i}$. The imputed values using CUT or POs are used as complete data in building CART and they lead to the same CART model as we implement CART algorithm with L_2 CUT or PO observed loss functions, respectively.

Remarks

At this point, we would like to comment that our methods are not limited to semiparametric forms in the sense that they do not depend on the widely used proportional hazard

assumptions. Furthermore, our methods serve as a natural extension of the regression tree to the interval censoring regime as the constructed loss functions reduce to the complete loss when there is no censoring.

2.2.3 Nonparametric Maximum Likelihood Estimation for Interval-Censored Data

Before moving on to the empirical studies, we give a brief review of the nonparametric maximum likelihood estimator (NPMLE) for interval-censored data. The NPMLE of a survivor function for right-censored data is given by the well-known Kaplan-Meier estimator (Kaplan and Meier, 1958). However, the NPMLE of a survivor function for interval-censored data does not have a closed-form in general and can only be determined using iterative algorithms.

Turnbull (1976) proposed a self-consistent algorithm which finds the limit of iterates obtained from the self-consistent equation (Efron, 1967). For interval-censored data where the failure time of subject i is T_i but we only observe an interval $(\mathfrak{L}_i, \mathfrak{R}_i]$, let $t_1, \ldots, t_m$ denote the unique ordered elements of $\{\mathfrak{L}_i, \mathfrak{R}_i : i = 1, \ldots, n\}$. For convenience we let $t_0 = 0$ and $t_{m+1} = \infty$ if necessary. We further define $p(t_j) = P(t_{j-1} < T \leq t_j) = S(t_{j-1}) - S(t_j)$ and $\alpha_{ij} = I(t_j \in (\mathfrak{L}_i, \mathfrak{R}_i])$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. It can be shown that the likelihood function only depends on $S(t)$ through values of $S(t_j)$, $j = 1, \ldots, m$. As a result, we are essentially maximizing the likelihood function over all discrete distributions that are constant between the increasing sequence of t_j, $j = 1, \ldots, m$. Turnbull (1976) further showed that the NPMLE of the distribution function increases in only a finite number of disjoint intervals, which are called the innermost intervals (Yu et al., 2000). The innermost intervals are defined as the set of disjoint intervals whose left points are in $\mathfrak{L}_i$, $i = 1, \ldots, n$ and right points are in $\mathfrak{R}_i$, $i = 1, \ldots, n$, and which contain no other members of $\mathfrak{L}_i$ or $\mathfrak{R}_i$ except the endpoints. The self-consistent algorithm is summarized as follows.

Initializing $\hat{p} = (\hat{p}(t_1), \ldots, \hat{p}(t_{m+1})) = \left(\frac{1}{m+1}, \ldots, \frac{1}{m+1}\right)$, we update the estimate at iteration r as

$$\hat{p}(t_j)^r = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}(t_j)^{r-1} \alpha_{ij}}{\sum_{k=1}^{m} \hat{p}(t_k)^{r-1} \alpha_{ik}},$$

for $j = 1, \ldots, m$. Repeating the process till convergence gives the self-consistent estimate of $p = (p(t_1), \ldots, p(t_{m+1}))$ as the solution to the self-consistent equations. Note that Turnbull's self-consistent MLE is not uniquely defined in the innermost intervals. We only know the amount of weight on the intervals but not the way the weight varies within the intervals. The algorithm can be viewed as an application of the EM algorithm using the Lagrange multiplier criterion from graph theory. Alternative algorithms for determining NPMLE include the ICM algorithm (Jongbloed, 1998) and the EM-ICM algorithm (Wellner and Zhan, 1997). See Section 3.4 of Sun (2006) for more details.

The NPMLE for current status data is different; see Section 2.5.1 for a brief review.

2.3 Simulation Studies

In this section, we empirically evaluate our proposed methods via simulation and compare them to *ad hoc* imputation and the conditional inference tree approach of Fu and Simonoff (2017). In Section 2.3.1, we describe the setting of the simulation study. In Section 2.3.2, we consider regression trees built on imputations of the failure times T and report their predictive performance and ability to recover the true data structure. In Section 2.3.3, we focus on regression trees built on the imputations of the failure status $I(T > u)$ at a landmark time and evaluate their predictive performance.

2.3.1 Simulation Set-up

We considered a sample size $n = 200$ with 500 replications. We generate $(W_1, W_2, W_3, W_4, W_5)$ from a multivariate normal distribution with zero mean and covariance matrix Σ of the form

$$\Sigma_1 = I_{5\times 5}, \qquad \text{or} \qquad \Sigma_2 = \begin{bmatrix} 1 & 0.9 & 0.9^2 & 0.9^3 & 0.9^4 \\ 0.9 & 1 & 0.9 & 0.9^2 & 0.9^3 \\ 0.9^2 & 0.9 & 1 & 0.9 & 0.9^2 \\ 0.9^3 & 0.9^2 & 0.9 & 1 & 0.9 \\ 0.9^4 & 0.9^3 & 0.9^2 & 0.9 & 1 \end{bmatrix},$$

which represent independent covariates and a highly correlated "autoregressive" dependence structure for the covariates, respectively. Based on these variables, we consider five covariates generated as follows:

1. $X_1 = I(W_1 < 0)$ (binary);

2. $X_2 = I(W_2 < Q_{0.25}) + 2I(Q_{0.25} \leq W_2 < Q_{0.5}) + 3I(Q_{0.5} \leq W_2 < Q_{0.75}) + 4I(Q_{0.75} \leq W_2)$, where Q_α is the α quantile of a standard normal distribution (ordinal);

3. $X_3 = I(W_3 < Q_{0.25}) + 2I(Q_{0.25} \leq W_3 < Q_{0.75}) + 3I(Q_{0.75} \leq W_3)$ (nominal);

4. $X_4 = e^{W_4}$ (continuous);

5. $X_5 = W_5$ (continuous).

We suppose that the data structure has a tree form with three terminal nodes and the failure time at each terminal node follows a Weibull distribution. Thus, we assume:

Node 1: $T \sim Weibull(\kappa_1, \lambda_1)$ if $X_2 \leq 2$;

Node 2: $T \sim Weibull(\kappa_2, \lambda_2)$ if $X_2 > 2$ and $X_4 > c$;

Node 3: $T \sim Weibull(\kappa_3, \lambda_3)$ if $X_2 > 2$ and $X_4 \leq c$.

See Appendix 2A at the end of the chapter for additional failure time distributions and data structures. We let $c = 1$ if covariates are generated independently ($\Sigma = \Sigma_1$) and $c = e^{0.611}$ if covariates are highly correlated ($\Sigma = \Sigma_2$) to guarantee the proportion of subjects falling into three terminal nodes to be 50%, 25%, and 25%, respectively.

Several constraints are imposed to determine the shape and scale parameters of the Weibull distributions including (i) the median of the marginal distribution of T is 5; (ii) the 0.9 quantile of the distribution of T at the second terminal node is 10; (iii) $\kappa_1 = \kappa_2$ and $\kappa_3 = 3$; (iv) the means of the three terminal nodes are set to be μ, $A\mu$ and $B\mu$, respectively, where A and B control the strength of the signal of the data. Table 2.1 provides a summary of node means according to choices of A and B. Four pairs of (A, B)

are selected to represent set-ups with different strength of the data signal. In the first set-up, all three node means are well-separated; in the second set-up, the first and second node means are close but the third node is far apart; in the third set-up, the second and third nodes are closer but the first node is far apart; in the fourth set-up, we consider a case in which the data signal is extremely weak so that all three node means are close.

Table 2.1: The choices of A and B and node means in four signal settings. Shape and scale parameters (κ, λ) of the Weibull distributions are displayed in the brackets following the node means.

Signal Settings	A	B	Mean of Node 1	Mean of Node 2	Mean of Node 3
1	2	4	3.58 (3.63, 3.97)	7.16 (3.63, 7.95)	14.32 (3, 16.04)
2	1.2	2.4	4.22 (1.43, 4.64)	5.06 (1.43, 5.57)	10.13 (3, 11.33)
3	2	2.5	3.78 (4.45, 4.15)	7.56 (4.45, 8.29)	9.45 (3, 10.58)
4	1.1	1.21	4.93 (1.67, 5.52)	5.42 (1.67, 6.07)	5.97 (3, 6.68)

We next generate the assessment times and allow subjects to have different numbers and timings of the assessments. For subject i, we generate $q_i \sim U(0.75, 0.99)$ and set the duration of follow-up τ_i as the $100q_i$th percentile of the marginal distribution of T. The number of assessments for subject i is then generated as $R_i \sim \text{Poi}(G_i \rho \tau_i)$, where $G_i \sim \Gamma(10, 10)$ and ρ was determined by setting the expected number of assessments $E(R_i) = 10$. The assessment times of subject i are then generated as R_i uniform random variables $u_{ir} \sim U(0, \tau_i)$, where $r = 1, \ldots, R_i$. We then let $u_{R_i+1} = \infty$ and $\Delta_{ir} = \sum_{r=1}^{R_i+1} I(T_i \in (u_{ir-1}, u_{ir}])$ and $(\mathfrak{L}_i, \mathfrak{R}_i] = (u_{ir-1}, u_{ir}]$, for $i = 1, \ldots, n$. This set-up addresses the heterogeneity of timings of the assessments across subjects.

2.3.2 Prediction of Failure Times

We propose regression trees for interval-censored failure time data based on the L_2 observed data loss functions using CUT in (2.4) and PO in (2.9). When predicting for failure times, the L_2 CUT observed data loss function can be implemented with L_2 complete

data regression trees based on CUT imputation and the imputed response is $\hat{Z}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$ in (2.5) with conditional mean (2.6); the L_2 PO observed data loss function can be implemented with PO imputation and the imputed response $\hat{Z}(O_i, X_i) = \hat{\theta}_{1i}$ given in (2.10).

We aim to compare the performance of our proposed regression tree based on response imputation for predicting failure times with the following benchmark methods:

1. (*Oracle*) the regression trees built on the uncensored failure times T_i;
2. (*M*) imputation using $(\mathfrak{L}_i + \mathfrak{R}_i)/2$, the midpoint of the interval $(\mathfrak{L}_i, \mathfrak{R}_i]$;
3. (*R*) imputation using $\mathfrak{R}_i$, the right endpoint of the interval $(\mathfrak{L}_i, \mathfrak{R}_i]$;
4. (*CIT*) conditional inference tree for interval-censored data proposed by Fu and Simonoff (2017). The influence function required to determine the splitting variable is chosen as the log-rank score (Pan, 1998)

$$\frac{\hat{S}(\mathfrak{L}_\mathfrak{i}) \log \hat{S}(\mathfrak{L}_\mathfrak{i}) - \hat{S}(\mathfrak{R}_\mathfrak{i}) \log \hat{S}(\mathfrak{R}_\mathfrak{i})}{\hat{S}(\mathfrak{L}_\mathfrak{i}) - \hat{S}(\mathfrak{R}_\mathfrak{i})},$$

   where $\hat{S}$ is the NPMLE of the survivor function constructed by the Turnbull's algorithm.

For response imputation, oracle tree, midpoint and right endpoint imputation, the regression trees are built using the `rpart` function from the `R` package `rpart` with the argument `method = "anova"`.

Estimation of the Survivor Function

The form of the CUT imputation $\hat{Z}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$ involves unknown conditional survivor function $S(\cdot|X)$, which is estimated semiparametrically under a Cox proportional hazard model (CUT_{Cox}) or nonparametrically using the conditional inference trees proposed by Fu and Simonoff (2017) (CUT_{Con}). The Cox model and conditional inference

tree are built upon the interval-censored data using the functions `ic.sp` and `ICtree` from the `R` packages `icenReg` and `LTRCtrees`, respectively. We also conduct CUT imputation using the marginal survivor function $S(\cdot)$ directly, which is estimated using the Turnbull's estimator (Turnbull, 1976) and realized using the `icfit` function from the `R` package `interval`. Linear interpolation is used to smooth the Turnbull's estimator of the marginal survivor function of T (Turnbull, 1976). In the case that the estimated survivor function does not decrease to zero, we fit a parametric tail implemented using the `survreg` function in the `R` package `survival` based on a hazard with a Weibull form.

Imputation based on the pseudo-observation (PO) $\hat{Z}(O_i, X_i) = \hat{\theta}_{1i}$ also utilizes the linearly smoothed Turnbull estimator with a parametric tail to estimate the marginal survivor function of T.

Evaluation Metrics for Prediction and Structure Recovery

Various evaluation metrics are considered to assess the performance of the methods through the test data set. The main attention is directed at prediction accuracy and the ability to recover the true tree structure.

The prediction error (PE) reflects the prediction accuracy and is defined as

$$PE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\mu_i - \hat{T}_i)^2 \ ,$$

where μ_i is the conditional expectation of T_i given X_i falling into a terminal node based on the true tree structure (i.e., $\lambda_1\Gamma(1+\frac{1}{\kappa})$, $\lambda_2\Gamma(1+\frac{1}{\kappa})$, $\lambda_3\Gamma(1+\frac{1}{3})$ in the three terminal nodes of the true tree, respectively) and $\hat{T}_i$ denotes the predicted failure time of subject i calculated based on the fitted tree.

The evaluation metrics for recover the true tree structure include:

Model Size: The average size of the fitted model (i.e., the number of terminal nodes of the fitted tree). In our setting, the closer this is to 3 the better the algorithm performs.

Number of Predictors (# Predictors): This is the average number of predictors (i.e., the

mean number of unique covariates the tree splits on). In our setting, the closer to 2, the better the performance.

Percent Correct (% Correct): This reflects the ability of the tree to split on the correct covariates, regardless of the splitting points and the order of splits. This is reported as the percentage of simulated samples for which the method split on both X_2 and X_4, so the higher the percentage, the better the performance.

Percent Without Noise (% w/o Noise): The ability to avoid noise variables. This is reported as the percentage of simulated samples for which the method did not inappropriately split on X_1, X_3, and X_5, so higher percentages correspond to better performance.

Results for Prediction Performance and Structure Recovery

Figure 2.3 shows the boxplots of PEs obtained under four signal settings as listed in Table 2.1 if the covariates are generated independently, or they are highly correlated. We compare our proposed CART algorithms based on *PO*, *CUT*, CUT_{Con} and CUT_{Cox} with the benchmark approaches listed above. Figure 2.3 contains eight subfigures. The subfigures in four rows correspond to the results from four signal settings from top to bottom; those in the left column correspond to the set-up with independent covariates, and those in the right column correspond to the scenario with the highly correlated covariates. In each subfigure, the order of the boxplots follows the oracle tree (*Oracle*), CART with imputed responses using *PO*, *CUT*, CUT_{Con}, CUT_{Cox}, conditional inference trees for interval-censored data (*CIT*) proposed by Fu and Simonoff (2017) and midpoint imputation (*M*). The results of the right endpoint imputation (*R*) are not shown as they have dramatically larger PEs as expected. As shown in Figure 2.3, our proposed CART algorithms perform consistently better than the conditional inference tree approach, especially when the covariates were highly correlated. In general, the CUT imputation methods provide closer performance to that of the oracle tree than the PO imputation. It is worth mentioning that when covariates are highly correlated, the CUT method based on the conditional inference tree estimation of the conditional survivor function $\hat{S}(\cdot|X)$ perform worse than the ones based on the Turnbull's and the Cox model estimations. This is because the conditional inference tree

approach does a poor job in prediction, and therefore provides an inaccurate estimation of the conditional survivor function. The CUT imputations based on estimations of the marginal survivor function using Turnbull's estimators and those of conditional survivor function using the Cox model are comparable under all scenarios. Mid-point imputation performs well across all settings, and our CUT imputations based on Turnbull's estimators and Cox model perform comparably or better than midpoint imputation. Finally, all methods behave comparably when the data signal is weaker in Settings 3 and 4. In such cases, the means of the terminal nodes are closer and therefore all the methods tend to fail to split and estimate the overall mean instead, which leads to low prediction errors. Moreover, note that when the data signal is weaker in Settings 3 and 4, the PEs are counter-intuitively smaller than those in settings 1 and 2, where the true node means are larger. Overall, we conclude that our methods are advantageous when making predictions of the interval-censored failure times compared to the conditional inference trees approach proposed by Fu and Simonoff (2017).

The performance of structure recovery is presented in Table 2.2 in terms of four evaluation metrics: *Model Size*, *# Predictors*, *% Correct* and *% w/o Noise*. As shown in Table 2.2, all methods perform reasonably well in Settings 1 when the three node means are far apart, which is the easiest setting for the methods to fully detect the true tree structure. In Setting 2, the first two terminal nodes have close mean values but they are under different primary splits. Therefore, most methods do a decent job of recovering true tree structure except that the conditional inference tree approach tends to miss some influential predictors and is only able to catch the correct covariates 64.2% of times when the covariates are independent and 19.8% of simulation trials when the covariates are highly correlated. When the second and third node means are close, the performance of all methods deteriorates as in Setting 3. When all the node means are close in Setting 4, none of the methods can effectively tell the nodes apart and tend to work with the entire training set. Across different methods, our CART algorithms based on various imputation responses perform better than the conditional inference tree approach in the sense that they fit closer to the size of the true tree, tend to pick up both influential predictors, have a larger chance to catch the correct covariate and avoid the noise covariates, and this advantage is more obvious when the covariates are highly correlated. In fact, our methods are comparable to

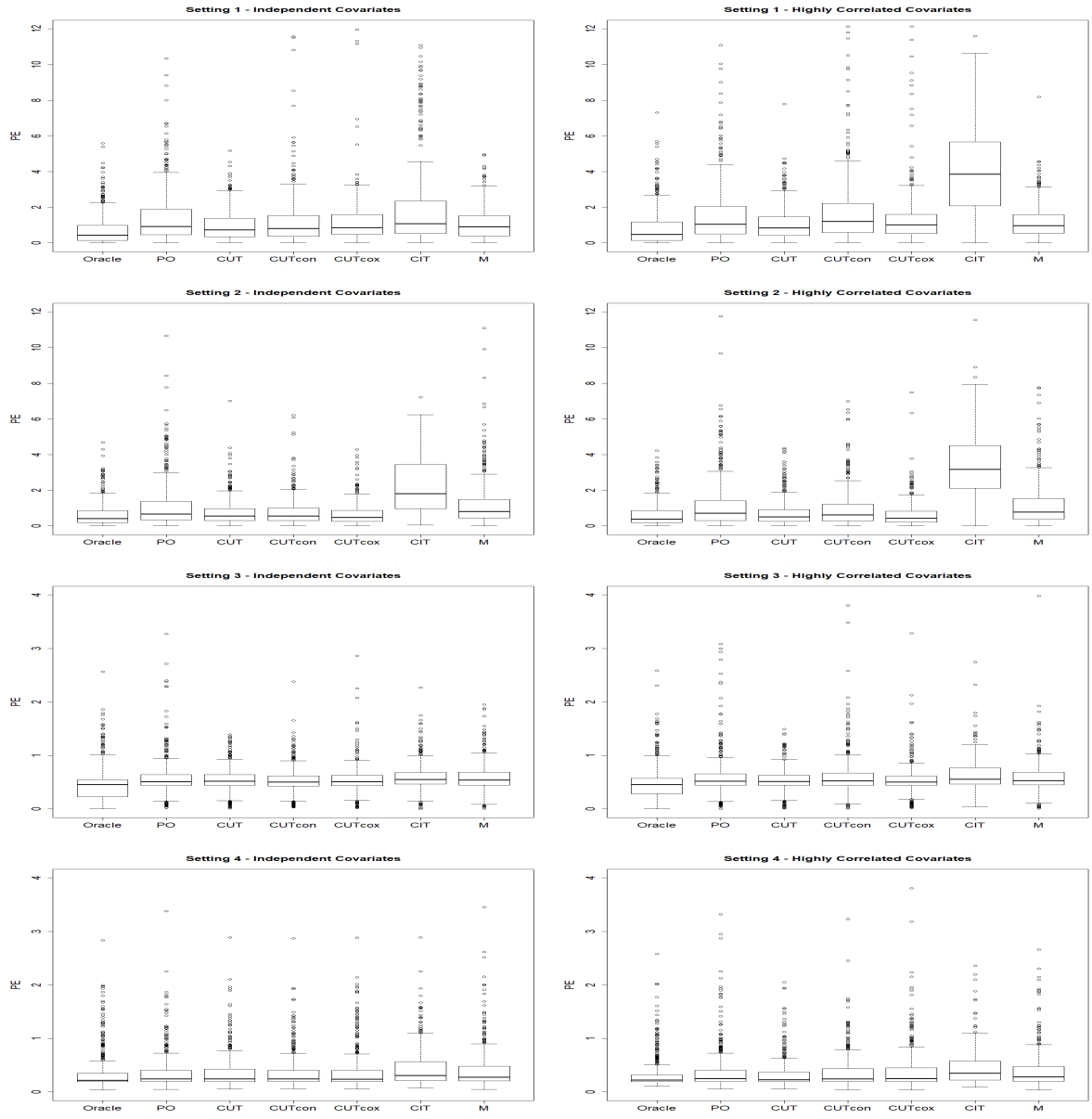

Figure 2.3: Prediction errors for predicting failure times under for four signal settings with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT$_{Con}$, CUT$_{Cox}$) for interval-censored failure time data and the benchmarks (Oracle, CIT, M).

the oracle tree except in the third setting.

Table 2.2: Structure recovery measures for four set-ups of strength of signals with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Oracle, CIT, M, R).

	Independent Covariates								Highly Correlated Covariates							
	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R
							Setting 1									
Model Size	3.294	3.374	3.268	3.370	3.468	3.038	3.248	3.262	3.376	3.456	3.288	3.662	3.414	3.598	3.374	3.350
# Predictors	2.136	2.162	2.122	2.156	2.166	1.992	2.110	2.106	2.164	2.158	2.096	2.296	2.148	2.478	2.116	2.116
% Correct	88.0	87.4	91.2	87.6	86.6	84.2	91.4	91.2	85.8	85.6	90.8	70.8	85.6	22.8	89.4	90.0
% w/o Noise	88.0	87.4	91.2	87.6	87.4	92.6	91.4	91.2	87.2	85.6	90.8	71.0	86.4	23.8	89.4	90.0
							Setting 2									
Model Size	3.302	3.296	3.314	3.280	3.320	2.838	3.226	3.236	3.374	3.324	3.290	3.450	3.310	3.222	3.234	3.240
# Predictors	2.130	2.108	2.128	2.112	2.134	1.808	2.100	2.094	2.132	2.108	2.116	2.190	2.124	2.144	2.096	2.092
% Correct	89.0	88.8	90.6	90.0	89.2	64.2	91.8	90.0	87.6	85.4	88.2	80.8	88.2	19.8	90.0	87.0
% w/o Noise	89.2	89.6	90.6	90.0	89.2	91.0	92.0	90.8	87.8	86.6	88.6	81.4	88.4	26.0	90.4	88.0
							Setting 3									
Model Size	3.012	2.576	2.522	2.528	2.608	2.242	2.545	2.547	2.964	2.564	2.518	2.746	2.516	2.452	2.538	2.540
# Predictors	1.772	1.446	1.388	1.436	1.436	1.226	1.411	1.414	1.728	1.430	1.388	1.534	1.348	1.410	1.388	1.384
% Correct	45.4	25.0	20.6	23.6	19.2	13.2	20.8	18.0	38.8	21.8	18.2	16.4	13.8	8.2	16.4	17.0
% w/o Noise	86.4	90.6	91.6	89.8	89.2	93.2	91.0	89.6	80.0	86.8	86.6	75.2	86.4	71.8	85.4	86.4
							Setting 4									
Model Size	1.550	1.502	1.642	1.606	1.670	1.192	1.558	1.592	1.478	1.516	1.488	1.676	1.794	1.210	1.476	1.550
# Predictors	0.434	0.396	0.514	0.498	0.514	0.192	0.452	0.452	0.370	0.412	0.398	0.526	0.652	0.212	0.390	0.434
% Correct	2.0	2.2	2.6	3.0	2.2	0.2	2.8	2.6	1.4	2.4	2.0	2.6	4.0	0.0	3.0	2.8
% w/o Noise	87.6	89.0	87.0	88.4	88.6	97.6	90.4	87.8	86.2	84.4	86.6	81.6	78.2	90.0	86.8	84.0

2.3.3 Prediction of Failure Status

In this subsection, we aim to compare the performance of our proposed regression trees based on response imputation for predicting failure status at a specific time point u (i.e., $I(T > u)$) with the following benchmark methods:

1. (*Oracle*) the regression trees built on the failure status $I(T_i > u)$;
2. (*M*) imputation using $I((\mathfrak{L}_i + \mathfrak{R}_i)/2 > u)$;
3. (*R*) imputation using $I(\mathfrak{R}_i > u)$;
4. (*CIT*) conditional inference tree for interval-censored data proposed by Fu and Simonoff (2017).

CUT imputed response is $\hat{Z}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$ in (2.5) with conditional mean (2.7) and the PO imputed response is $\hat{Z}(O_i, X_i) = \hat{\theta}_{1i}$ given in (2.11). Time point u is chosen as the 0.25, 0.50 and 0.75 quantiles of the marginal distribution of T.

Evaluation Metrics for Prediction Performance

Prediction errors ($PE_{survivor}$) are computed to measure difference between the true and estimated survival probabilities at time point u, defined by

$$PE_{survivor} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[p_i(u) - \hat{p}_i(u) \right]^2 ,$$

where

$$\begin{aligned} p_i(u) &= \sum_{k=1}^{K} I(X_i \in \mathcal{X}_k) P(T_i > u | X_i \in \mathcal{X}_k), \\ \hat{p}_i(u) &= \sum_{k=1}^{K^*} I(X_i \in \mathcal{X}_k^*) \frac{\sum_{j=1}^{n} I(X_j \in \mathcal{X}_k^*) I(\hat{T}_j > u)}{\sum_{\ell=1}^{n} I(X_\ell \in \mathcal{X}_k^*)}, \end{aligned}$$

$\hat{T}_j$ represent the imputed value of the j subject, $j = 1, \ldots, n$, $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$ is the true partition of the covariate space according to the true tree structure, $\{\mathcal{X}_1^*, \ldots, \mathcal{X}_{K^*}^*\}$ is a

partition of the covariate space following the fitted tree, and K^* is the number of terminal nodes in the fitted tree.

Results for Prediction Performance

The boxplots of prediction errors for predicting failure status at the 0.25, 0.50, and 0.75 quantiles of the marginal distribution of the failure time are shown in Figures 2.4, 2.5, and 2.6, respectively. In each figure, the subfigures and boxplots are organized in the same way as those in Figure 2.3. As shown in Figures 2.4, 2.5, and 2.6, the conditional inference tree approach is relatively not stable across assessment times and signal settings. For instance, in Figure 2.6, the conditional inference tree approach leads to significantly larger prediction errors than our methods for Settings 1 and 2 with highly correlated covariates. Our proposed imputation methods are either comparable or better than the conditional inference tree approach in most settings, although the conditional inference tree performs the best in Setting 2 in Figure 2.4 by around 0.005 in terms of the median. While the regression trees are fitted based on the true failure status or the corresponding imputations, the conditional inference trees are fitted with the left and right endpoints of the censoring intervals. The fact that the conditional inference trees take in data of different nature may explain their counter-intuitive superior performance (even to the oracle method) in such a setting. Among our proposed imputation methods, the CUT imputations are either comparable or better than the PO imputation. Overall, CUT imputation based on the Turnbull's estimation of the marginal survivor function $\hat{S}$ and the Cox model of the conditional survivor function $\hat{S}(t|X)$ are consistently superior in predicting the failure status across assessment times and simulation set-ups.

2.4 Analysis of Data from a Study of Axial Disease

In this section, we apply our methods to analyse data from a study aiming to identify genetic markers associated with the development of axial disease and predict the status of axial disease in patients with psoriatic arthritis.

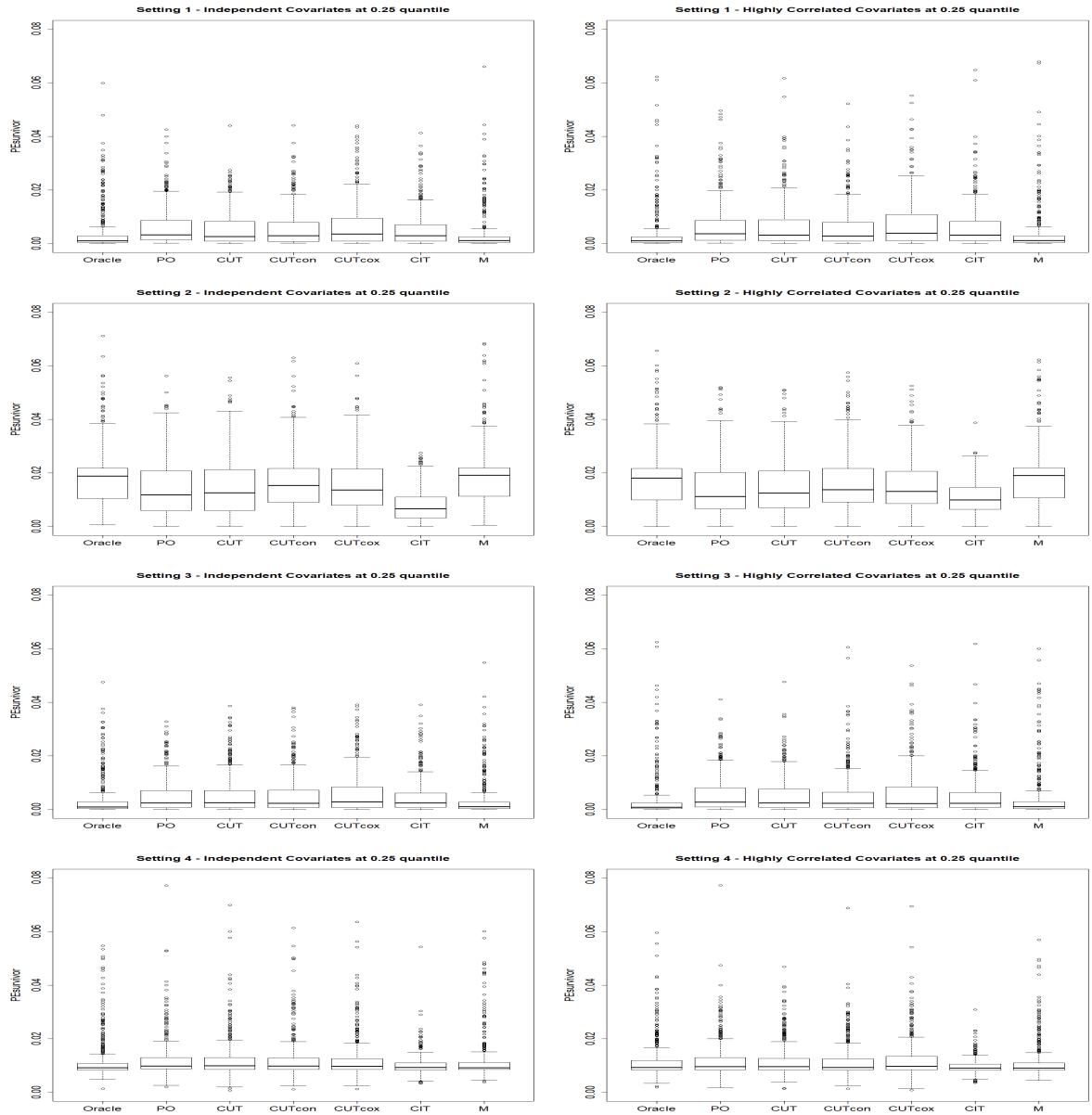

Figure 2.4: Prediction errors for predicting failure status at 0.25 quantile of the marginal distribution of the failure time for four signal settings with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Oracle, CIT, M).

In Setting 2, CIT outperforms the oracle method, which may be explained by the different nature of input data.

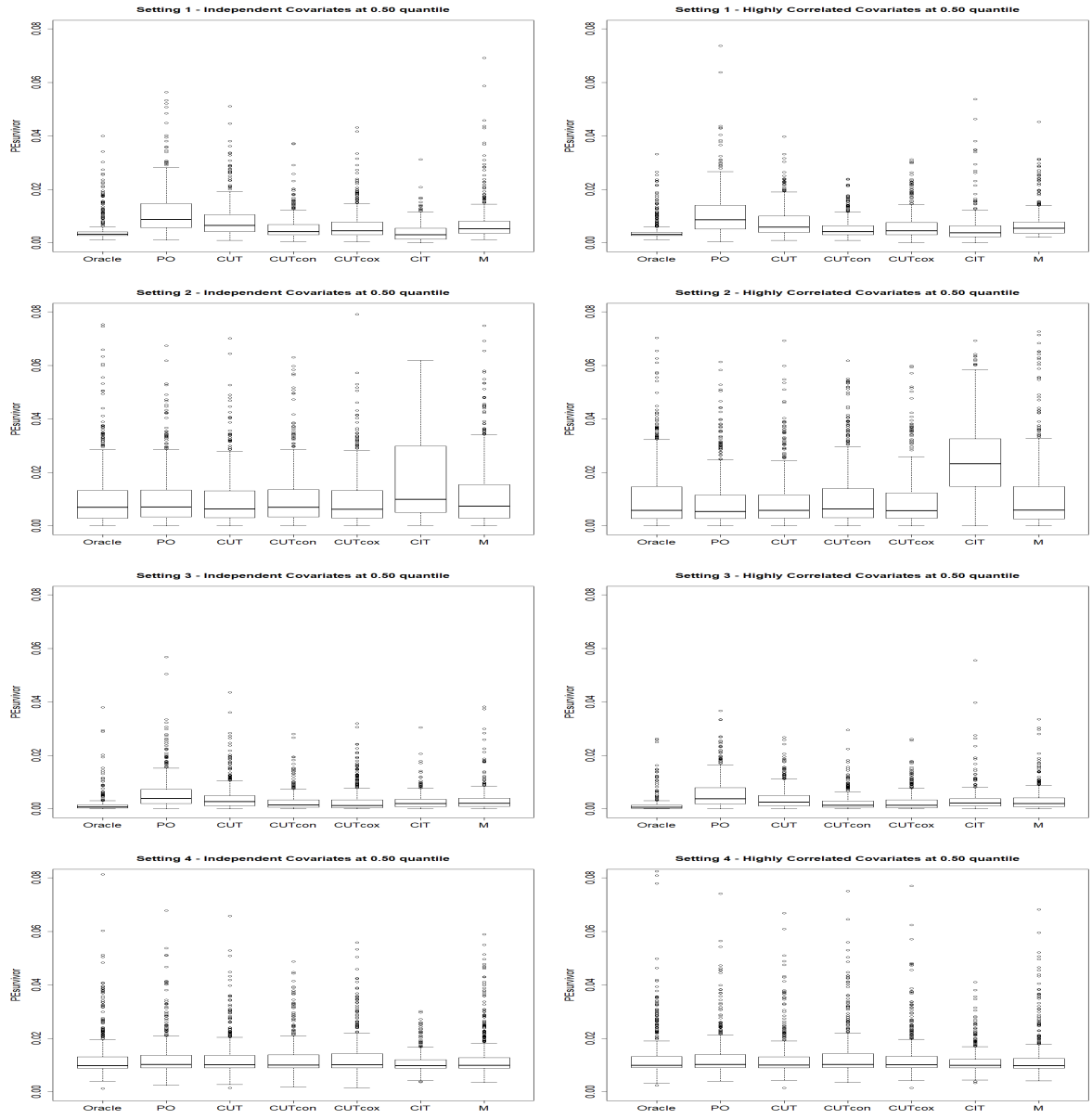

Figure 2.5: Prediction errors for predicting failure status at 0.50 quantile of the marginal distribution of the failure time for four signal settings with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT$_{Con}$, CUT$_{Cox}$) for interval-censored failure time data and the benchmarks (Oracle, CIT, M).

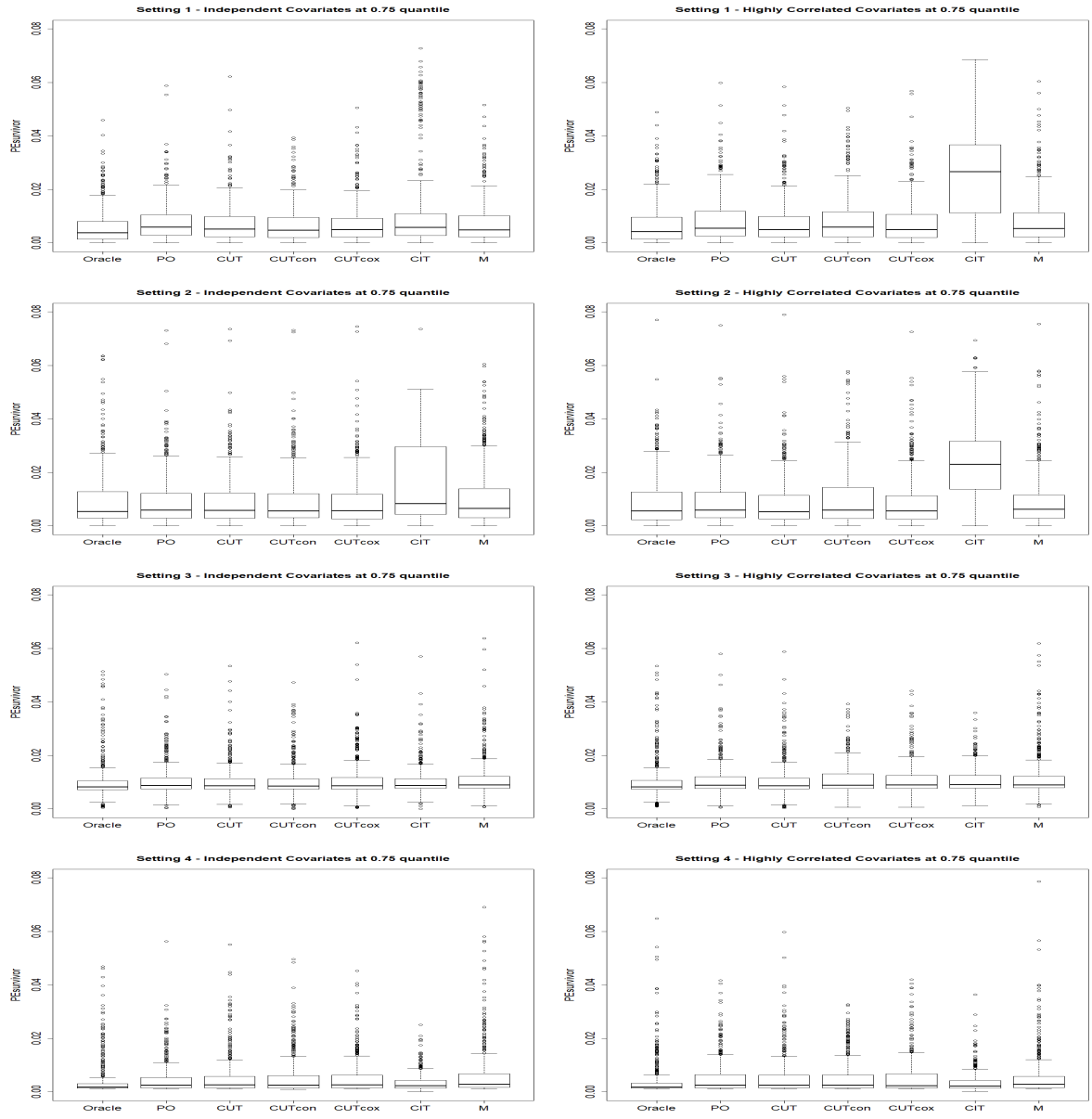

Figure 2.6: Prediction errors for predicting failure status at 0.75 quantile of the marginal distribution of the failure time under four signal settings with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Oracle, CIT, M).

2.4.1 Data Description

The University of Toronto Psoriatic Arthritis Registry maintains a large cohort of patients with psoriatic arthritis, an autoimmune disease that features joint inflammation and damage along with skin involvement (Gladman and Chandran, 2011). A subset of the full registry comprised of 1022 patients includes data on Human Leukocyte Antigen (HLA) markers. A particular concern in these patients is the development of spinal involvement, referred to as axial disease, which has a serious impact on functional ability and mobility. The presence of axial disease is recorded if both the left and the right sacroiliac (SI) joints score at least grade 2 damage on the New York criteria or one of the SI joints has at least grade 3 damage (Bennett and Burch, 1967). Data on HLA biomarkers are available; we code these as 1 representing the biomarker to be present and 0 representing the absence of the biomarker. In addition, we have the following explanatory variables:

age.ps: Age at diagnosis of psoriasis (years);
age.psa: Age at diagnosis of psoriatic arthritis (years);
sex.female: Sex of the patient, with 1 for female and 0 for male;
fmhx.ps: Family history of psoriasis with 1 for yes and 0 for no;
fmhx.psa: Family history of psoriatic arthritis with 1 for yes and 0 for no.

The data set further contains the following data related to the interval-censored time to axial involvement.

Axial.Ltime: The age of a patient at his or her last radiographic assessment, which shows that the disease has not developed;
Axial.Rtime: The age of a patient at his or her first radiographic assessment which shows that the disease has already developed;
Axial.status: An indicator on whether the subject is right-censored (disease has not developed at the last assessment).

The failure time of interest is the number of years a patient takes to develop the axial disease after the diagnosis of psoriatic arthritis. Therefore, the left and right endpoints of the censoring interval of our interest are *Axial.Ltime* - *age.psa* and *Axial.Rtime* - *age.psa*, respectively. Approximately 62% of the patients were not observed to develop the axial disease, so they had right-censored responses.

2.4.2 Identification of Potential Influential Predictors

We aim to identify potential influential predictors for the development of axial disease in this subsection. According to Chandran et al. (2010b), the HLA biomarker *B27* affects the development of the axial disease. All explanatory covariates and biomarkers described in the preceding subsection are put into the learning algorithms and we are interested in whether the algorithms are able to successfully identify *B27* as an influential predictor.

Trees for predicting failure times are built using the CART algorithm based on PO imputation (PO), CUT imputation (CUT), midpoint imputation (M), and the conditional inference tree (CIT) approach by Fu and Simonoff (2017). The PO and CUT imputed values are computed based on Turnbull's estimates of the marginal survivor function. Using ten-fold cross-validation, it is observed that the CARTs are sensitive to how the data set is partitioned into cross-validation sets and return different sizes of the optimal subtrees. Therefore, we repeat the cross-validation steps for all the CART algorithms by creating 200 different partitions of the cross-validation sets, going through the cross-validation procedure 200 times and recording the size of the optimal subtree each time. The final tree is selected as one with the most frequent size of the selected subtrees in the 200 repetitions.

We summarize the results in Figure 2.7, which provides the fitted tree structures, fitted survival curves in the terminal nodes of the fitted trees, and the barplots showing the frequency of the size of the selected optimal subtrees across 200 cross-validations of PO, CUT, and M trees. The final trees built using CUT, M, and CIT (not shown) are the same with the first split on *sex.female* and the second split on the biomarker *B27* in the male subgroup, but the final tree built using PO has only one split on *age.psa*. The fitted survival curves at the terminal nodes of fitted trees are quite separated, suggesting all the learning algorithms do a decent job of stratifying the patients into distinct risk groups. The barplot of the PO method contains eight bars and the frequencies 2 and 6 are competitive, however, the barplot of CUT contains five bars and the frequency 2 is dominating, suggesting that the CARTs based on PO imputation are more sensitive to different partitions of the data set than those based on CUT imputation. Overall speaking, the CARTs based on CUT is able to identify influential predictor for failure time of interest as well as CIT and midpoint imputation. The CARTs based on PO is less stable and fails

to identify the influential predictor.

Trees for predicting failure status in years 5, 10 and 15 are built using PO and CUT methods. As shown in Figure 2.8, both PO and CUT trees at five years split on *B27* first and then *sex.female* in the *B27* present subgroup; PO and CUT trees at ten years and CUT tree at fifteen years split on *sex.female* first and then *B27* in the male subgroup, which also agrees with the PO, M, and CIT trees when predicting for failure time; PO at fifteen years further splits on *age.psa* in the *B27* non-present subgroup, which leads to two terminal nodes with close predicted values and seems unnecessary. All models identify *B27* as well as *sex.female* as influential predictors.

Overall, the primary significant predictors are *sex.female* and the biomarker *B27*. It takes longer for the female patients to develop the axial disease since the diagnosis of psoriatic arthritis than the male patients. In addition, it takes shorter for patients with biomarkers *B27* present to develop the axial disease since the diagnosis of psoriatic arthritis.

2.4.3 Predicting Failure Status

In this section, we evaluate the predictive accuracy for the failure status of the axial disease after ten years of psoriatic arthritis.

The data set is divided into a training data set and a test data set (70% and 30%, respectively, according to the 70/30 rule) such that they have approximately the same proportion of right-censored patients. Fitting the tree algorithms on the training data set, we compared the prediction performance of PO and CUT to M and CIT evaluated using the test data set. While the responses in the test data set are still interval-censored, a patient is sure to be axial disease-free 10 years later from the diagnosis of psoriatic arthritis if the left endpoint of the censoring interval is greater than 10 and is axial disease-present if the right endpoint of the censoring interval is smaller than 10. If an interval from the test data set contains 10, the patient is at risk and the probability of being disease-free $P(T > 10|\mathfrak{L} < T \leq \mathfrak{R})$ is estimated using the Turnbull's estimated survival curve obtained from the test data set.

Here we calculate the prediction error by averaging the difference between the disease

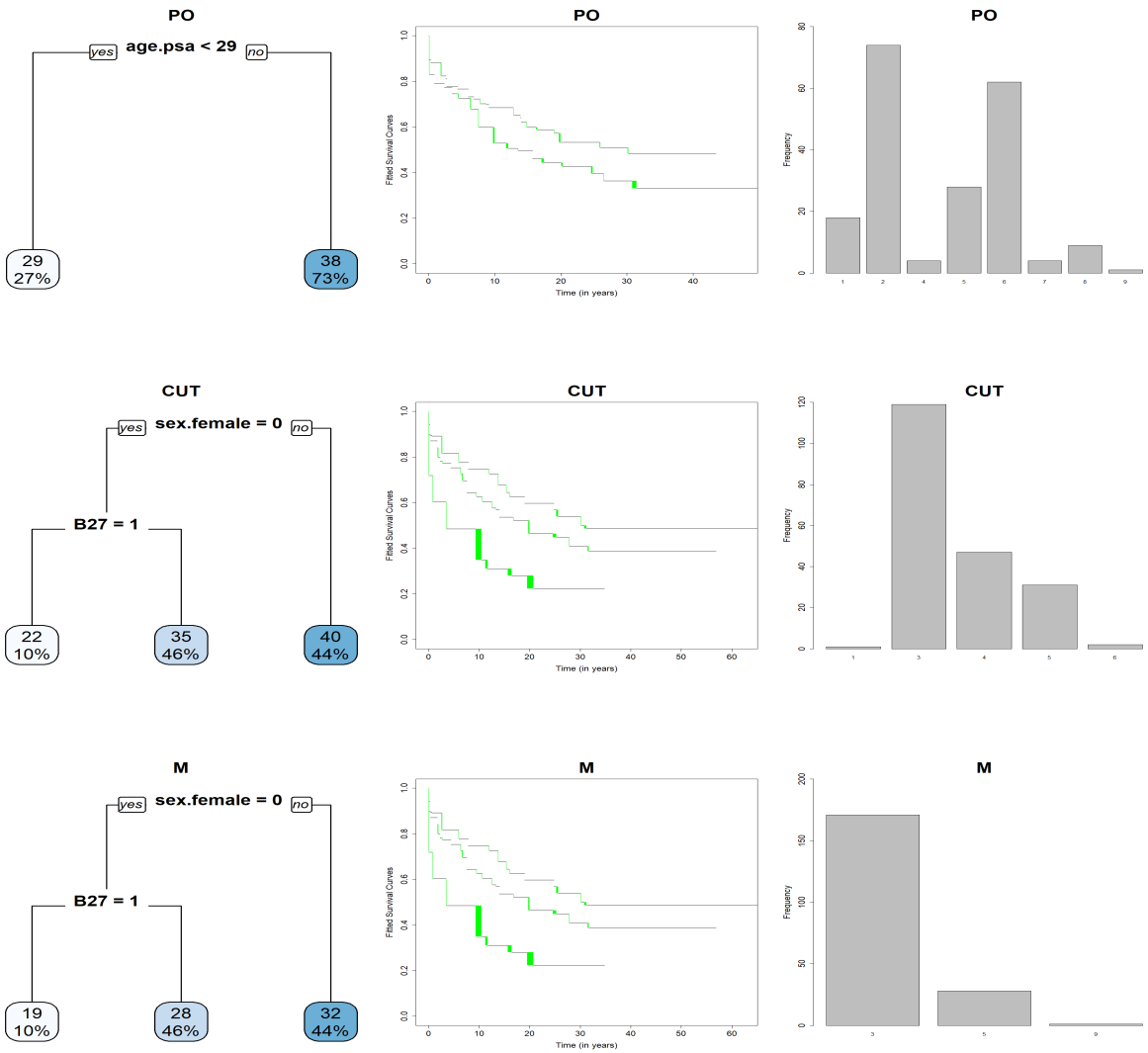

Figure 2.7: The fitted tree structures, fitted survival curves in the terminal nodes of the fitted trees, and the barplots showing the frequency of the size of selected optimal subtrees across 200 cross-validations of PO, CUT and M trees.

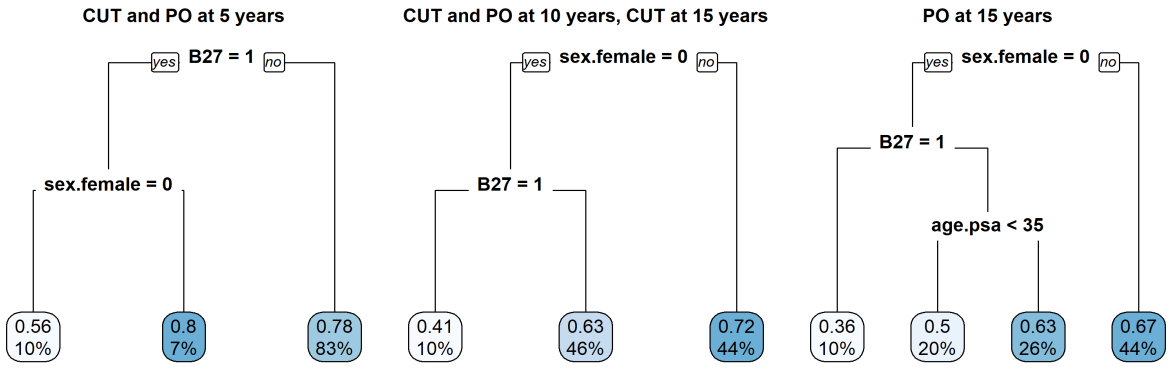

Figure 2.8: Fitted trees from disease status after 5, 10, and 15 years of diagnosis of psoriatic arthritis using PO and CUT methods.

status predicted by the tree algorithms and the disease status obtained from the test data set,

$$PE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left\{ \Delta_i (Z_i - \hat{Z}_i)^2 + (1 - \Delta_i) E[(Z_i - \hat{Z}_i)^2 | \mathfrak{L}_i < T_i \leq \mathfrak{R}_i, X_i] \right\}$$

where $Z_i = I(T_i > 10)$, $\hat{Z}_i$ is the predicted disease status of patient i after ten years of psoriatic arthritis obtained by rounding the predicted probability from the tree models to zero or one, $\Delta_i = I(\mathfrak{L}_i > 10) + I(\mathfrak{R}_i \leq 10)$ indicates if the disease status of patient i after ten years of psoriatic arthritis is observed in the test data set. The prediction errors of PO, CUT, M and CIT are 0.352, 0.324, 0.324, 0.341, respectively. The results are consistent with our impression in simulation studies that CUT and M are the most advantageous in terms of prediction accuracy compared to PO and CIT approaches.

2.5 Survival Trees for Current Status Data

Here we adapt the strategies discussed in Section 2.2 to provide a general framework for implementing the CART algorithm to build survival trees while accommodating a current status observation scheme of the response time of interest. As a special type of interval-

censored data, current status data give rises to non-trivial extensions of the ideas in Section 2.2.

2.5.1 Current Status Data

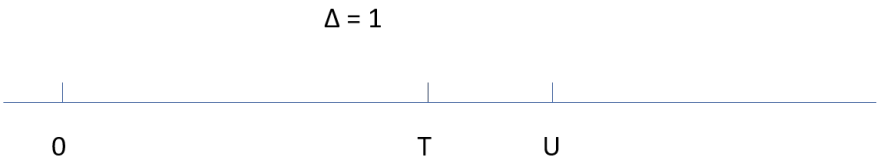

Figure 2.9: An illustration of current status data.

Current status data, also known as type I interval-censored data, arise if there is a single random examination time U ($U > 0$), and it is only known whether or not the event time T exceeds U; we let $\Delta = I(T \leq U)$, so the observed data for a particular individual is $O = (U, \Delta)$. Recall that conditional on the covariates that are controlled for in the analysis, the event time is assumed to be independent of the examination time as in Cook and Lawless (2019). With a sample of observations on n independent processes, we denote the *complete data* as $\mathcal{D} = \{Z_i, X_i : i = 1, ..., n\}$ and the *observed data* as $\mathcal{O} = \{O_i, X_i : i = 1, \ldots, n\}$. Unlike general type R interval censoring, the NPMLE of the survivor function of current status data has a closed form, which we briefly review as follows.

Let $U_{(j)}$, $j = 1, \ldots, m$ denote the unique ordered elements of $\{0, U_1, \ldots, U_n\}$, let $n_j = \sum_{i=1}^{n} I(U_i = U_{(j)})$ denote the number of individuals who are assessed at $U_{(j)}$, and let $r_j = \sum_{i=1}^{n} \Delta_i I(U_i = U_{(j)})$ denote the number of individuals assessed and found to have failed at the examination $U_{(j)}$, $j = 1, \ldots, m$. The likelihood function

$$L(S(\cdot)) = \prod_{i=1}^{n} S(U_i)^{1-\Delta_i}[1 - S(U_i)]^{\Delta_i}$$

can be written as

$$\prod_{j=1}^{m} S(U_{(j)})^{n_j - r_j}[1 - S(U_{(j)})]^{r_j} = \prod_{j=1}^{m} F(U_{(j)})^{r_j}[1 - F(U_{(j)})]^{n_j - r_j}$$

where $F(t) = 1 - S(t)$. Therefore, the likelihood function only depends on the survivor function at the inspection times. With the constraint of $F(U_{(1)}) \leq \ldots \leq F(U_{(m)})$, the optimization problem is equivalent to an isotonic regression (Robertson et al., 1988) problem involving data $\{r_1/n_1, \ldots, r_m/n_m\}$ with weights $\{n_1, \ldots, n_m\}$. According to the maximum-minimum formula for isotonic regression, the NPMLE is

$$\hat{F}(U_{(j)}) = \max_{v_1 \leq j} \min_{v_2 \geq j} \frac{\sum_{l=v_1}^{v_2} r_l}{\sum_{l=v_1}^{v_2} n_l}.$$

So the NPMLE of the survivor function $S(t) = P(T \geq t)$ has a closed form $\hat{S}(t) = 1 - \hat{F}(t)$ for current status data. In practice, $\hat{S}(t)$ can be computed empirically via the pooled adjacent violators algorithm (PAVA) for isotonic regression (Barlow et al., 1972).

2.5.2 The Observed Data Loss Functions Based on Censoring Unbiased Transformations and Pseudo-Observations

We focus on the case in which the interest lies in the failure time itself (i.e., $Z = T$). The complete data loss function is $L(\mathcal{D}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} L(T_i, \Psi(X_i))$ and $\mathcal{R}(\Psi) = E[L(T, \Psi(X))]$ is the complete data risk. As shown in Section 2.2.2, we construct observed data loss functions which are unbiased or consistent for the complete data risk based on CUTs and POs. Following the construction of the L_2 observed data loss via CUT, we need to find a suitable function $\mathscr{Y}_j^*(O, X)$ in (2.4) that are the CUT for T^j, $j = 1, 2$. Since

$$E(T^j|X) = \sum_{\delta=0}^{1} E(T^j|\Delta = \delta, X) P(\Delta = \delta|X),$$

the CUT of T^j can be constructed as

$$\mathscr{Y}_j^*(O, X) = \sum_{\delta=0}^{1} I(\Delta = \delta) E(T^j|\Delta = \delta, X), \tag{2.13}$$

where,

$$E(T^j|\Delta = 1, X) = -\frac{\int_0^U t^j dS(t|X)}{1 - S(U|X)},$$

and

$$E(T^j|\Delta = 0, X) = -\frac{\int_U^\infty t^j dS(t|X)}{S(U|X)}. \tag{2.14}$$

Thus, $\mathscr{Y}_j^*(O, X)$ is a CUT of T^j since it has the same conditional expectation as T^j given covariates X.

Similarly, following the construction of the L_2 observed data loss via PO, if $\hat{\theta}_j$ in equation (2.9) is an estimator of $E(T^j)$ and $\hat{\theta}_j^{(-i)}$ is the corresponding leave-one-out estimator, then $\hat{\theta}_{ji} = n\hat{\theta}_j - (n-1)\hat{\theta}_j^{(-i)}$ is the ith PO for $E(T^j)$, $j = 1, 2$. We estimate $E(T^j) = -\int_0^\infty t^j dS(t)$ by replacing $S(t)$ with an estimate so that the POs for $E(T^j)$ can be written as

$$\hat{\theta}_{ji} = -n \int_0^\infty t^j d\hat{S}(t) + (n-1) \int_0^\infty t^j d\hat{S}^{(-i)}(t) \ , \tag{2.15}$$

where $\hat{S}(\cdot)$ is NPMLE of the survivor function $S(\cdot)$ and $\hat{S}^{(-i)}(\cdot)$ is the corresponding leave-one-out estimator excluding data of individual i.

With the same argument as in Section 2.2.2, we adapt the idea of response imputation to current status data. With $Z = T$, we can make the imputed loss function (2.12) equivalent to the L_2 CUT loss function (2.4) by letting $\hat{T}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$ given in (2.13) or to the L_2 PO loss function (2.9) by letting $\hat{T}(O_i, X_i) = \hat{\theta}_{1i}$ given in (2.15). The imputed values using CUTs or POs are used as complete data in building CART, and they lead to the same CART model as we implement the CART algorithm with L_2 CUT or PO observed loss functions, respectively.

2.5.3 Simulation Studies

We now assess the performance of our methods empirically by how well they recover the correct tree structure and predict failure times. Our proposed methods are compared via simulations to the oracle tree built using uncensored event times, methods based on *ad hoc* approaches, and the conditional inference tree approach of Fu and Simonoff (2017), which was originally designed for interval-censored data.

Data Generation

We considered a sample size $n = 500$ with 500 replications. The covariates and the data structure of a tree form follow from Section 2.3.1, with the means of the three terminal nodes set to be 5μ, 4μ and 2μ, respectively. The node means are found as 7.48, 5.99, and 2.97, respectively.

We next generate the examination times. We let τ denote a maximum time of interest beyond which no assessments will be scheduled, and τ is set as the 95th quantile of the marginal distribution of T. We further let $U^* \sim \Gamma(\alpha, \beta)$ with Gamma cumulative distribution function G. The examination time is set as $U = \min(U^*, \tau)$. We let $\rho = P(T < U)$ represent the proportion of individuals who fail at their examinations in the population. By adjusting the rate and scale parameters of the Gamma distribution the variability of the examination times vary. For a specified proportion of individuals who fail at their examinations in the population, we have

$$\begin{aligned}\rho &= P(T < U, T < \tau, U^* < \tau) + P(T < U, T < \tau, U^* > \tau) \\ &= \int_0^\tau P(T \leq u | U = u) dG(u; \alpha, \beta) + P(T < \tau)(1 - G(\tau)).\end{aligned}$$

Hence, this set-up not only addresses the heterogeneity of timings of the examination times across subjects but also allows us to investigate the effect of informative assessments and the variability of the inspection times on performance by choosing different values of ρ and $\text{Var}(U^*)$, respectively. For each specified ρ and $\text{Var}(U^*)$, we can solve for α and β accordingly, followed by generating the assessments from the gamma distribution with an upper limit τ. Table 2.3 provides a summary of parameter configuration across various choices of ρ and $\text{Var}(U^*)$.

Methods for Event Time Prediction

We propose regression trees for current status data based on the L_2 observed data loss functions using CUT in (2.4) and PO in (2.9). When predicting for event times, the L_2 CUT observed data loss function can be implemented with L_2 complete data regression trees based on CUT imputation, and the imputed response is $\hat{T}(O_i, X_i) = \mathscr{Y}_1^*(O_i, X_i)$ in (2.13)

Table 2.3: Parameter configuration for the distribution of the examination times.

$\text{Var}(U^*)$	ρ	α	β	$P(U^* > \tau)$	$\text{Var}(U^*)$	ρ	α	β	$P(U^* > \tau)$
1	0.30	11.98	3.46	2.33×10^{-9}	4	0.30	3.16	0.89	1.13×10^{-3}
	0.50	26.08	5.11	1.07×10^{-8}		0.50	7.13	1.34	2.20×10^{-3}
	0.70	55.31	7.44	3.60×10^{-6}		0.70	14.64	1.91	1.26×10^{-2}

with conditional mean (2.14); the L_2 PO observed data loss function can be implemented with PO imputation, and the imputed response is $\hat{T}(O_i, X_i) = \hat{\theta}_{1i}$ given in (2.15). Imputation based on the pseudo-observation (PO) $\hat{T}(O_i, X_i) = \hat{\theta}_{1i}$ utilizes the linearly smoothed nonparametric estimator of the marginal survivor function of T obtained from the PAVA via the `gpava` function in the `R` package `isotone`. The unknown conditional survivor function $S(\cdot|X)$ involved in the CUT imputation is estimated semiparametrically under a Cox proportional hazard model (CUT_{Cox}) or nonparametrically using the conditional inference trees (CUT_{Con}) as in Section 2.3.2. The Cox model and conditional inference tree are implemented by expressing the current status data in the form of interval-censored data and using the functions `ic.sp` and `ICtree` from the `R` packages `icenReg` and `LTRCtrees`, respectively, since they are originally developed for interval-censored data. When using the package `LTRCtrees`, we can either use the conditional survivor function estimates obtained from the package `LTRCtrees` (CUT_{Con}) or directly estimate the conditional survivor functions by using the PAVA in each terminal node of the fitted conditional inference tree (CUT_{ConP}).

We compare the performance of our proposed regression trees based on response imputation for predicting event times with the following benchmark methods listed in Section 2.3.2:

1. Oracle trees (O).
2. Right imputation (R): When $T \leq U$, the imputed value takes U. If $T > U$, the imputed value is chosen as the time point at which the marginal survivor function estimate decreases to zero; in case it does not decrease to zero, it is chosen as the

time point at which the marginal survivor function estimate reduces to the minimal value.

3. Midpoint imputation (M): When $T \leq U$, the imputed value takes $U/2$. When $T > U$, the imputed value takes the average of U and the right-imputed value;

4. Conditional inference trees (CIT).

Prediction and Structure Recovery

We adopt the evaluation metrics for prediction and structure recovery in Section 2.3.2. Figures 2.10 and 2.11 summarize the performance of the proposed CART algorithm based on PO, CUT_{Cox}, CUT_{ConP}, and CUT_{Con} compared to the benchmark approaches listed above. The variability of inspection times is larger in Figure 2.10 ($\text{Var}(U^*) = 4$) and smaller in Figure 2.11 ($\text{Var}(U^*) = 1$). Within each figure, the set-ups with independent covariates are presented in the left column and the set-ups with highly correlated covariates are presented in the right column. The proportion of individuals who fail at their examination times are 30%, 50%, and 70% down the columns. Our proposed regression trees based on CUT_{ConP} had the best performance across the set-ups. The regression trees based on CUT_{Cox}, CUT_{Con} and PO also outperformed the conditional inference trees in most set-ups. All the methods perform worse than the oracle tree in predicting event times to a reasonable extent considering how much less information the current status data contains than the complete data. Furthermore, most tree algorithms deteriorate when the covariates are highly correlated and ρ is smaller. When the inspection times are less informative, the conditional inference trees are less stable and may produce extremely large PEs if some terminal nodes are full of right-censored individuals as shown in the first row of Figure 2.11. The regression trees based on midpoint imputation have small PEs as illustrated in Figure 2.10. However, they lead to considerable larger PEs than the other methods when assessments are less informative ($\rho = 0.3$ and 0.5) and in the meanwhile, inspection times are less variant ($\text{Var}(U^*) = 1$) in Figure 2.11. Finally, the PEs based on right imputation are not shown in Figures 2.10 and 2.11 as the PEs are too large to fit in the figures of the current scale.

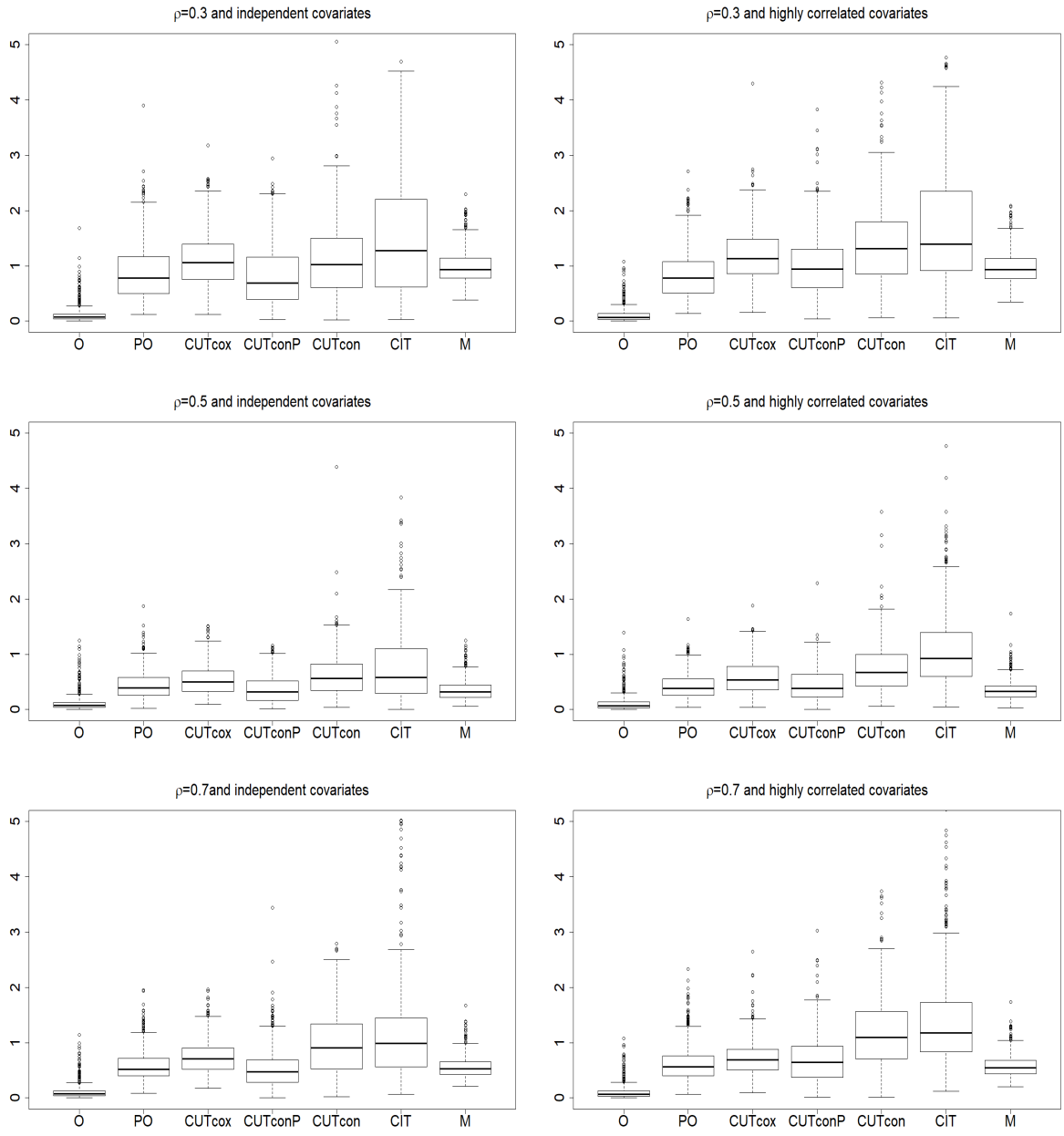

Figure 2.10: Prediction errors for predicting event times comparing proposed survival tree algorithms (PO, CUT_{Cox}, CUT_{Con}, CUT_{ConP}) for current status data and the benchmarks (O, CIT, M); $\text{Var}(U^*) = 4$.

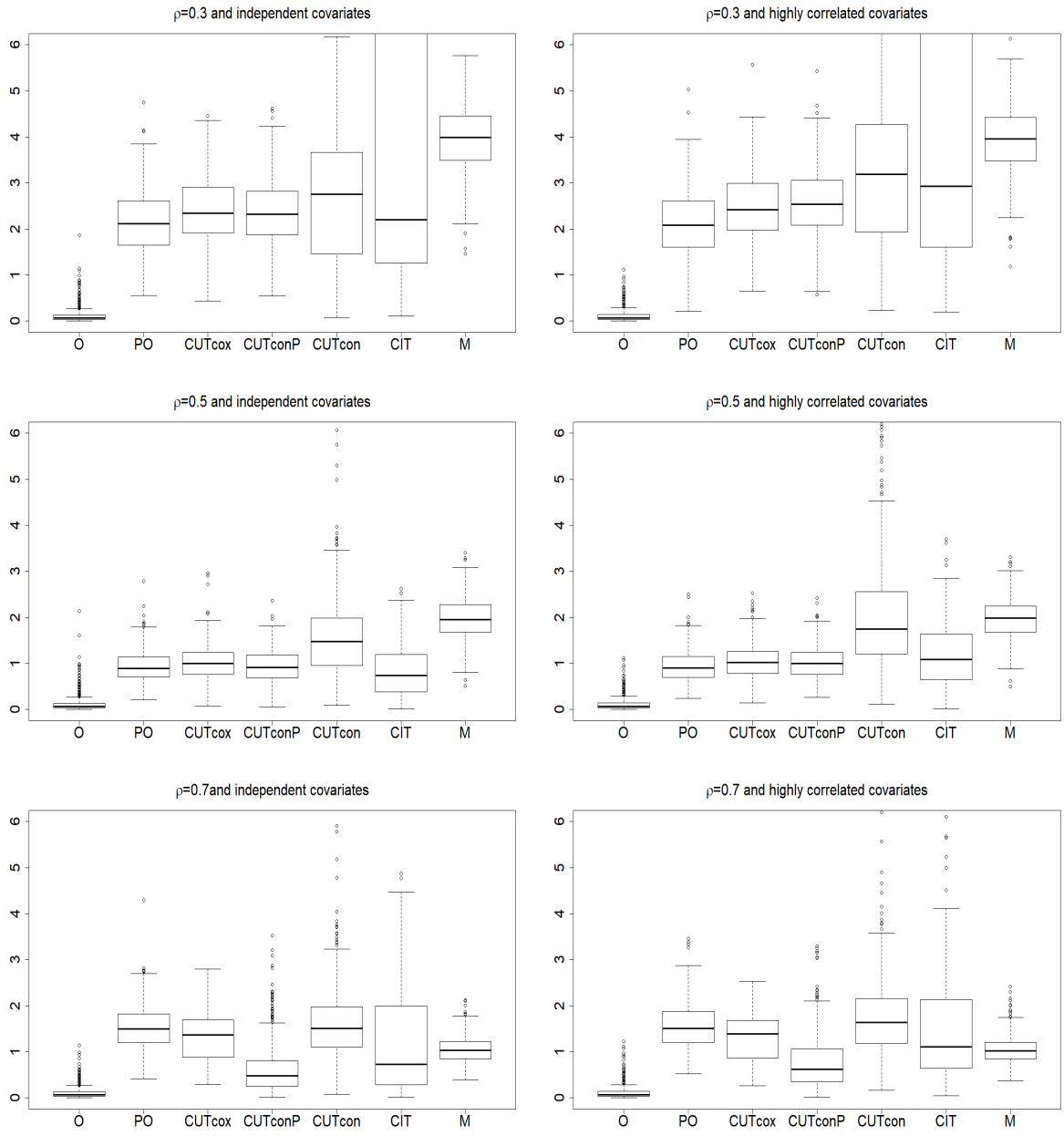

Figure 2.11: Prediction errors for predicting event times comparing proposed survival tree algorithms (PO, CUT_{Cox}, CUT_{Con}, CUT_{ConP}) for current status data and the benchmarks (O, CIT, M); $\text{Var}(U^*) = 1$.

Table 2.4 summarizes the structure recovery performance of the proposed survival tree models and the benchmarks. The top half of the table contains results of the set-ups having assessments with higher variability ($\text{Var}(U^*) = 4$), and the bottom half displays results of the set-ups having assessments with lower variability ($\text{Var}(U^*) = 1$). The left half of the table shows the set-ups with independent covariates, and the right half shows those with highly correlated covariates. The proportion of individuals found to fail at the examination times are 30%, 50%, and 70% down the columns. The survival trees based on PO recover the underlying tree structure well, and their results are comparable to those of the oracle trees in most set-ups, which is valuable as the current status data contains much less information than the complete data. When the covariates are independent, the conditional inference trees perform well. However, the conditional inference trees frequently fail to recover the underlying tree structure when the covariates are highly correlated as they tend to pick up some noise variables and build larger trees than the true tree structure. The survival trees based on CUT_{ConP} and CUT_{Con} perform comparably well to or slightly worse than the oracle trees when covariates are independent. Nevertheless, their performance deteriorates in the set-ups with highly correlated covariates as their conditional survivor functions are estimated by the conditional inference trees and hence, undermined by the compromised performance. It is noteworthy that the survival trees based on CUT_{ConP} and CUT_{Con} still perform consistently better than the conditional inference trees. The regression trees based on CUT_{Cox} do not recover the underlying tree structure as well as other survival trees, and they suffer from higher computational costs. Finally, both midpoint imputation and right imputation recover the underlying tree structure very well across all set-ups.

2.6 Discussion

In this chapter, we propose strategies to construct observed data loss functions for the interval-censored failure times to use in place of complete data loss in the CART algorithm and implement them using the imputation techniques. We build L_2 complete data regression trees based on imputed responses using CUT and PO, which enable us to reveal influential covariates and effectively predict the failure times and the failure status. As

Table 2.4: Structure recovery measures comparing proposed survival trees algorithms (PO, CUT_{Cox}, CUT_{Con}, CUT_{ConP}) for current status data and the benchmarks (O, CIT, M, R) under various settings.

	Independent Covariates								Highly Correlated Covariates							
	O	PO	CUT_{cox}	CUT_{conP}	CUT_{con}	CIT	M	R	O	PO	CUT_{cox}	CUT_{conP}	CUT_{con}	CIT	M	R
							$\text{Var}(U^*) = 4$ and $\rho = 0.3$									
Model Size	3.11	3.12	6.02	3.23	3.18	2.90	3.13	3.15	3.09	3.17	6.43	3.82	3.64	3.46	3.11	3.13
# Predictors	2.05	2.03	2.56	2.08	2.05	1.85	2.05	2.03	2.03	2.05	2.85	2.46	2.35	2.35	2.03	2.03
% Correct	95.6	91.0	59.4	83.2	85.0	67.0	93.4	88.6	97.0	89.4	41.0	29.8	35.4	17.8	92.8	88.4
% w/o Noise	95.6	94.2	60.0	87.6	89.8	89.8	94.4	92.6	97.0	91.0	41.2	30.8	37.2	20.4	94.2	91.8
							$\text{Var}(U^*) = 4$ and $\rho = 0.5$									
Model Size	3.11	3.09	3.85	3.25	3.23	3.08	3.10	3.11	3.09	3.06	4.11	3.63	3.55	3.95	3.15	3.14
# Predictors	2.05	2.04	2.23	2.13	2.11	2.02	2.03	2.05	2.03	2.02	2.37	2.34	2.32	2.80	2.06	2.06
% Correct	95.6	95.6	81.4	88.2	89.8	82.8	96.8	95.6	97.0	97.4	70.2	65.2	67.4	15.8	94.6	93.4
% w/o Noise	95.6	96.0	81.6	88.6	90.2	90.4	96.8	95.6	97.0	97.6	70.4	65.2	67.4	16.4	94.6	93.4
							$\text{Var}(U^*) = 4$ and $\rho = 0.7$									
Model Size	3.11	3.12	3.41	3.19	3.20	2.91	3.14	3.10	3.09	3.15	3.53	3.78	3.71	3.73	3.09	3.10
# Predictors	2.05	2.02	2.12	2.07	2.08	1.84	2.06	2.03	2.03	2.01	2.20	2.43	2.38	2.59	2.03	2.01
% Correct	95.6	90.2	80.2	86.8	86.8	69.2	94.4	92.8	97.0	86.0	75.6	41.8	48.8	19.0	94.0	90.2
% w/o Noise	95.6	94.4	85.8	90.4	89.4	92.0	94.6	95.0	97.0	92.6	78.0	43.4	49.8	21.6	95.4	93.8
							$\text{Var}(U^*) = 1$ and $\rho = 0.3$									
Model Size	3.11	3.07	4.24	3.24	3.32	3.03	3.15	3.18	3.09	3.07	4.51	3.89	3.96	3.75	3.12	3.19
# Predictors	2.05	2.01	2.26	2.11	2.03	1.98	2.06	2.07	2.03	2.02	2.48	2.52	2.42	2.61	2.05	2.07
% Correct	95.6	92.8	80.8	88.8	86.4	78.2	94.4	92.4	97.0	93.4	63.6	43.2	39.4	15.2	95.2	92.0
% w/o Noise	95.6	96.2	80.8	89.4	92.2	89.8	94.6	93.0	97.0	95.2	63.6	43.4	44.6	16.2	95.2	92.0
							$\text{Var}(U^*) = 1$ and $\rho = 0.5$									
Model Size	3.11	3.07	3.39	3.20	3.22	3.12	3.11	3.10	3.09	3.12	3.51	3.73	3.80	4.12	3.14	3.13
# Predictors	2.05	2.02	2.11	2.08	2.06	2.06	2.05	2.03	2.03	2.04	2.19	2.45	2.39	2.91	2.05	2.05
% Correct	95.6	96.8	91.0	92.6	91.2	88.4	95.8	97.0	97.0	95.0	83.2	58.4	59.2	14.2	94.6	94.8
% w/o Noise	95.6	97.2	91.0	92.6	93.2	91.4	95.8	97.0	97.0	95.4	83.2	58.4	61.5	14.4	94.6	94.8
							$\text{Var}(U^*) = 1$ and $\rho = 0.7$									
Model Size	3.11	3.00	3.33	3.20	3.31	2.97	3.10	3.11	3.09	3.01	3.25	3.68	3.64	3.71	3.18	3.11
# Predictors	2.05	1.89	2.10	2.05	2.09	1.89	2.04	2.02	2.03	1.85	2.09	2.40	2.34	2.57	2.05	2.01
% Correct	95.6	73.0	82.8	85.8	85.6	71.2	91.8	88.6	97.0	69.4	84.4	44.0	53.2	21.0	92.0	87.0
% w/o Noise	95.6	92.4	87.4	90.4	89.0	90.8	94.0	93.6	97.0	91.6	86.8	47.2	54.8	22.6	93.4	93.0

shown in the simulation studies and data analysis, our methods can predict more accurately than the conditional inference tree approach across settings of distinct strength of the data signals, especially when the covariates are highly correlated. The survival tree methods for interval-censored data designed by Yin and Anderson (2002) involved an exponential tree model which made strong parametric assumptions on the failure time distributions and a non-parametric model which heavily relied on large sample size and low dropout rate. Without any report on the prediction performance, approaches in Yin and Anderson (2002) were not included in our simulations for comparison. Our results advocate the CUT imputation based on the marginal survivor function estimated from Turnbull's algorithm and the conditional survivor function estimated from the Cox model to fit regression trees on interval-censored failure time data, since they are less computationally intensive compared to the PO method and more accurate than the conditional inference tree approach in both predicting failure times and failure status. In Appendix 2A, we report further simulation studies aiming to assess the performance of the methods in settings with various failure time distributions and underlying structures. Assuming another three types of distributions for the failure times under the terminal nodes gives similar results to those obtained in Section 2.3. However, additional investigations show that when the underlying structure does not have a tree form, the CUT imputation based on the marginal survivor function estimated from Turnbull's algorithm becomes unstable in prediction. We also find that in such settings, the conditional inference tree is less affected by the dependence structure of the covariates, while the performance of the regression trees based on traditional imputations deteriorates. Overall, we recommend the CUT imputation based on the conditional survivor function estimated from the Cox model for its best performance throughout the settings.

Additionally, we adopt the proposed strategies to construct survival trees when the available data arise from an independent current status observation scheme. Able to reveal influential covariates and make predictions, our methods can predict the event times more accurately than the conditional inference tree approach across a variety of assessment time models in terms of variance of the assessment times, the proportion of right-censored individuals, and dependence structures among the covariates in simulation studies. Our methods are shown to perform particularly well in recovering the underlying tree structures.

When aiming to fit regression trees based on current status data, we recommend the use of the PO imputation approach and the CUT imputation based on the conditional survivor function estimated using the pooled adjacent violators algorithm for each of the terminal nodes of the conditional inference trees.

This work is based on the assumption that the assessments are independent of the failure times given the covariates, but given the covariates are selected in a data-driven way, this is essentially equivalent to a completely independent inspection time. If there is concern about covariate dependent inspection time model, one can consider the use of inverse density-weighted loss functions to ensure consistent estimation of the complete data loss function (Zhu et al., 2017).

It is well-known that ensemble methods have advantages over single prediction models in terms of stability. With this in mind, Yao et al. (2019) extended the work from Fu and Simonoff (2017) to explore the use of ensemble methods based on the conditional inference survival forest. In ongoing work, we are adapting these regression tree algorithms for both interval-censored data and current status data to accommodate ensemble methods such as random forests. See Section 5.2.1 for a brief plan. Another potential research area is to investigate additional means of estimating the marginal or conditional survivor functions, for example, using the nonparametric Bayesian accelerated failure times approach. Finally, we note that the equivalence between the constructed observed data loss and the imputed loss only holds under the L_2 loss function in the CART algorithm - an extension of these methods to deal with different loss functions is an important area of future research, too.

Chapter 3

Adaptive Two-Phase Designs: Some Results on Robustness and Efficiency

3.1 Introduction

Large cohort studies often involve the creation of biobanks in which serum or tissue samples are collected from individuals upon recruitment and stored. After obtaining follow-up data on individuals in the cohort, scientific questions often arise about the association between a biomarker and a particular response. Two-phase studies aim to use follow-up data and inexpensive baseline covariate data to develop sampling strategies for the creation of a sub-sample of individuals with complete data on the biomarker. The goal of such a design is typically to obtain a more efficient estimator for the parameter of interest while meeting budgetary constraints which preclude evaluation of the biomarker in all individuals (Breslow and Cain, 1988; Breslow and Chatterjee, 1999; Lawless et al., 1999). One such study in rheumatology involves a registry of patients with psoriatic arthritis (PsA) at the University of Toronto PsA clinic (Gladman and Chandran, 2011). Investigators aim to study the association between the bio-marker matrix metalloproteinase 3 (MMP-3) and progression in joint damage. Blood and urine samples for those patients are stored in a bio-bank, so it is possible to assay the bio-specimens for some, but not all, individuals (Chandran et al., 2010a). A traditional marker of inflammatory disease, the erythrocyte

sedimentation rate (ESR), is relatively inexpensive and available for the whole cohort; we aim to exploit the auxiliary data to help select the phase II sub-sample for measurements of MMP-3.

The followup response and the inexpensive baseline data from the full cohort form the phase I sample, while the phase II sub-sample is comprised of individuals chosen from the phase I sample to have their biospecimens assayed. This phase II sub-sampling is typically carried out following stratification of the phase I sample. Traditional methods of phase II sub-sampling include simple random sampling, proportional sampling in which individuals are sampled proportionally to the stratum sizes, and balanced sampling in which equal numbers of individuals are chosen from each stratum. Lawless et al. (1999) and Breslow and Chatterjee (1999) described methods of estimation and inference based on the observed data likelihood which requires modelling the distribution of the biomarker. In contrast, the mean score estimating functions developed by Reilly and Pepe (1995) relax the need for these distributional assumptions at a cost of some efficiency. Inverse probability weighted estimating equations (Robins and Rotnitzky, 1995; Tsiatis, 2006) offer another approach in which contributions are made to the primary estimating function from individuals with known biomarker data, with a selection model used to address the fact that this is a biased sub-sample. In order to exploit the efficiency advantages of likelihood inference and relax the need to specify nuisance models, Chatterjee et al. (2003) proposed an innovative semiparametric maximum likelihood approach in which the distributions for the biomarker are estimated nonparametrically for different classes of individuals. More recently, Scott and Wild (2011) built upon Lawless et al. (1999) and proposed use of conditional likelihood while examining its relationship with several semiparametric efficient methods.

McIssac and Cook (2015) proposed an adaptive two-phase design that divides the phase II sub-sampling into phase IIA and IIB stages; the process naturally extends to multiple stages of phase II sub-sampling. The adaptive approximately optimal design exploits one of the established sampling schemes to select individuals into the phase IIA sub-sample and this data are used to obtain preliminary parameter estimates with which the asymptotically optimal sampling scheme can be obtained for the phase IIB sub-sampling. McIssac and Cook (2015) consider use of inverse probability weighted estimating functions and focused on minimizing the variance of the estimator from the phase IIB sub-sample; the impact of

the phase IIA selection model does not arise in their framework. In more recent work, Chen and Lumley (2020) consider multiwave response-dependent sampling to optimize design-based estimators with informative priors specified in order to minimize over-sampling of uninformative strata at any stage.

We extend the work of McIssac and Cook (2015) in several directions. We first redefine the optimal phase IIB selection criteria by considering the asymptotic distribution of the estimator of interest which is obtained upon completion of phase II sub-sampling; the precision of this estimator is influenced by the design used in the phase IIA pilot stage and so optimal phase IIB selection models may differ according to the sub-sampling scheme used in phase IIA. Second, while McIssac and Cook (2015) focused on estimating functions, we consider adaptive two-phase designs for maximum likelihood (Breslow and Chatterjee, 1999; Lawless et al., 1999), semiparametric maximum likelihood (Chatterjee et al., 2003), and conditional likelihood (Scott and Wild, 2011) estimation; the latter two are particularly appealing when the expensive covariate is continuous. Third, we extend the investigation to consider the case in which a surrogate of the biomarker of interest is available in the phase I sample. Fourth, we carry out a detailed empirical investigation regarding the efficiency and robustness of estimators under different strategies to design and approaches to analyse.

The remainder of the chapter is structured as follows. We introduce our methods and outline the procedure of an adaptive two-phase response-dependent sampling design for both discrete and continuous covariates in Section 3.2. In Section 3.3, we investigate the finite sample performance of estimators from the different designs in a variety of settings with a focus on efficiency and robustness. We also report on a study framed within a research project aiming to assess the relationship between an expensive biomarker MMP-3 and the development of new joint damage in psoriatic arthritis. In Section 3.4, we report on further investigations in the framework of the surrogate value problem. Section 3.5 concludes with some general remarks.

3.2 Adaptive Two-Phase Designs

3.2.1 Notation

Let Y denote the response, X_1 an expensive exposure variable of interest, X_2 a vector of discrete auxiliary covariates, and let $X = (X_1, X_2')'$. We consider a response model $f(Y|X;\beta)$ indexed by a vector of parameters β. The interest lies in estimation and inference regarding a particular component of β which, in what follows, we take to be the component β_1, corresponding to the coefficient of X_1 in a regression model. We let $G(X;\gamma)$ be the joint distribution of the covariate vector indexed by γ, where $g_1(X_1|X_2;\gamma_1)$ denotes the conditional probability density (mass) function of X_1 given X_2 when X_1 is continuous (discrete), and X_2 has the probability mass function $g_2(X_2;\gamma_2)$; we let $\gamma = (\gamma_1', \gamma_2')'$. Overall, the joint model of the response and covariates is

$$f(Y, X;\vartheta) = f(Y|X;\beta)g_1(X_1|X_2;\gamma_1)g_2(X_2;\gamma_2), \tag{3.1}$$

where $\vartheta = (\beta', \gamma')'$; we also let $\vartheta_1 = (\beta', \gamma_1')'$. In some settings, X_2 may be a surrogate for X_1 in which case $Y \perp X_2|X_1$; we consider this problem in Section 3.4.

In two-phase designs, individuals in a phase I sample of size n provide information on their response and inexpensive auxiliary covariates and we denote this data by $\{Y_i, X_{i2} : i = 1, \ldots, n\}$. A sub-sample of M individuals is then selected and the expensive covariate X_1 is measured in these individuals. The sub-sampling of individuals into the phase II sub-sample is governed by the selection model with sampling probabilities $\pi_i(Y_i, X_{i2};\psi)$ indexed by the parameter ψ, $i = 1, \ldots, n$, In this framework, the covariate X_1 is therefore missing at random (MAR) in the phase I sample (Little and Rubin, 2002).

In the regime of an adaptive two-phase design, we let A, π_A, M_A, and ψ_A denote the phase IIA selection indicator, selection probability, sub-sample size, and the parameter of the selection model, respectively. We adopt the letters B, C, and so on, to denote the corresponding features in the subsequent stages of the phase II process, conditioning on individuals not having been selected in previous stages; see Appendix 3A for details. Finally, we assume that ψ and ϑ are functionally independent.

3.2.2 Overview of Adaptive Two-Phase Designs and Stratification

For a given approach to analysis, the optimal selection model governs a phase II sub-sampling scheme which yields an estimator of β_1 with the minimum asymptotic variance among those possible with the same expected phase II sub-sample size. Such designs, however, require knowledge of unknown parameters. The adaptive two-phase approach involves a conventional sampling scheme for a fraction of the individuals to be selected in phase II to form a phase IIA sub-sample; this can be used to obtain a preliminary estimate of ϑ with which to approximate the optimal phase IIB selection model.

Let M_A ($M_A < M$) denote the size of the phase IIA sub-sample chosen based on *i)* simple random sampling, *ii)* proportional sampling, or *iii)* balanced sampling schemes of which the latter two depend on the stratification. We let $C(Y_i, X_{i2})$ record the stratum for individual i, where $C(\cdot)$ is a coarse mapping of (Y_i, X'_{i2}) to strata labeled as $1, \ldots, S$, $i = 1, \ldots, n$. The mapping $C(\cdot)$ determines the nature of the stratification of the phase I sample and hence how conventional proportional or balanced (Breslow and Cain, 1988) phase II sub-sampling schemes might be employed. We let $V_{is} = I(C(Y_i, X_{i2}) = s)$ for $s = 1, \ldots, S$ and $V_i = (V_{i1}, \ldots, V_{iS})'$ denote an $S \times 1$ vector identifying the stratum to which individual i belongs and define a stratum-dependent selection model so that X_{i1} is MAR (Little and Rubin, 2002). Let $\pi_{iA} = P(A_i = 1|V_i; \psi_A)$. Note that each of strategies *i)*, *ii)* and *iii)* lead to a particular choice of π_{iA} and hence ψ_A. Specifically, strategies *i)* and *ii)* induce identical π_{iA} for all individuals in phase I sample, except that *ii)* ensures to select the same proportion of individuals from each stratum. Strategy *iii)*, on the other hand, induces different π_{iA} across strata such that an equal number of individuals are selected from each stratum, i.e., inversely proportional to the stratum sizes. The preliminary parameter estimates obtained based on the phase IIA sample are denoted by $\tilde{\vartheta}$.

Following phase IIA sub-sampling, a total of $N_B = n - M_A$ individuals remain eligible for phase IIB sub-sampling. Subject to budgetary constraints, we aim to select M_B individuals according to a phase IIB sampling scheme to minimize the variability of the estimator for the parameter of interest from the whole phase II sub-sample and to approximate the

optimal selection model with $\pi_{iB} = P(B_i = 1|V_i, A_i = 0; \psi_B)$ while satisfying the sampling constraints. The approximately optimal ψ_B is defined as the one that minimizes

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \tilde{\vartheta}, \psi] + \lambda \left[E(B|A = 0; \tilde{\vartheta}, \psi) - \frac{E(M_B)}{E(N_B)} \right], \tag{3.2}$$

where $\psi = (\psi'_A, \psi'_B)'$, λ denotes the Lagrange multiplier, and the phase IIA parameter estimates $\tilde{\vartheta}$ are substituted into the expressions. The objective function and the constraint are functions of ψ_B, which here is a one-to-one function of π_B. See Figure 3.1 for an illustration of locating the minimal asymptotic variance of estimator of parameter of interest at the optimal ψ_B.

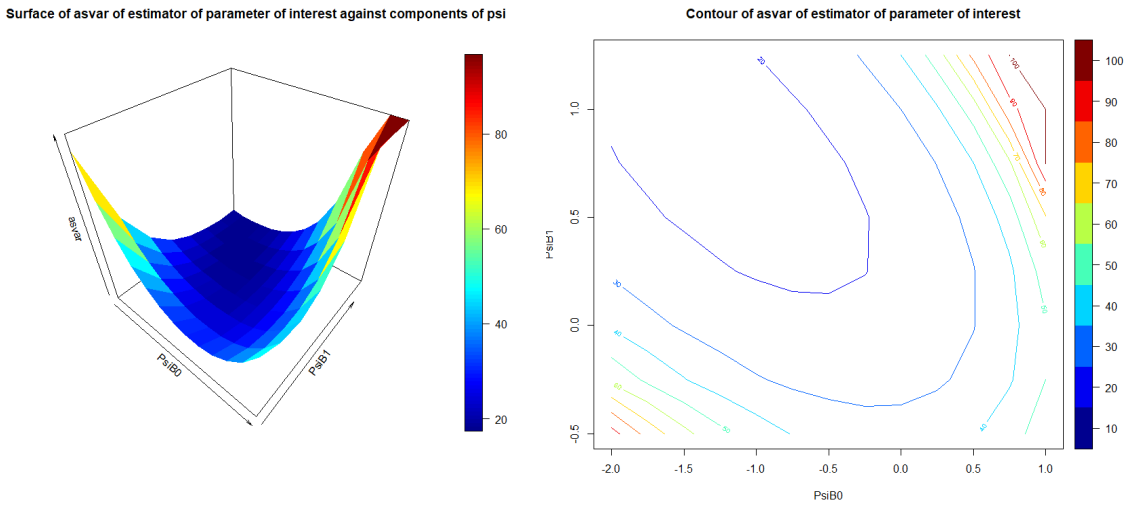

Figure 3.1: Surface (left) and contour (right) of asymptotic variance of the estimator of parameter of interest generated as a function of ψ_B. Minimal asymptotic variance locates at the optimal ψ_B.

3.2.3 Design and Analysis of Adaptive Two-Phase Studies

We next consider different frameworks for analysis along with the corresponding approaches to conduct adaptive two-phase designs. Note that when X_1 is continuous, strategies that avoid the need to model the nuisance distribution $g_1(X_1|X_2; \gamma_1)$ are particularly desirable.

Maximum Likelihood

Following the construction of a phase IIA sub-sample, the preliminary parameter estimates are obtained. If ψ and ϑ are functionally independent, we can restrict attention to the observed data likelihood

$$L(\vartheta) = \prod_{i=1}^{n} [f(Y_i|X_i;\beta)g_1(X_{i1}|X_{i2};\gamma_1)g_2(X_{i2};\gamma_2)]^{A_i} \left[\int f(Y_i|x_1, X_{i2};\beta)g_1(x_1|X_{i2};\gamma_1)g_2(X_{i2};\gamma_2)dx_1\right]^{1-A_i},$$

and hence, the log-likelihood $l(\vartheta) = \sum_{i=1}^{n} l_i(\vartheta)$ where

$$l_i(\vartheta) = A_i[\log f(Y_i|X_i;\beta) + \log g_1(X_{i1}|X_{i2};\gamma_1)] + (1-A_i)\left[\log \int f(Y_i|x_1, X_{i2};\beta)g_1(x_1|X_{i2};\gamma_1)dx_1\right] + \log g_2(X_{i2};\gamma_2).$$

The phase IIA estimate $\tilde{\vartheta}$ is obtained by solving

$$\frac{\partial l(\vartheta)}{\partial \vartheta} = \sum_{i=1}^{n} \frac{\partial l_i(\vartheta)}{\partial \vartheta} = \left(\frac{\partial l_i(\vartheta)}{\partial \beta'}, \frac{\partial l_i(\vartheta)}{\partial \gamma_1'}, \frac{\partial l_i(\vartheta)}{\partial \gamma_2'}\right)' = 0,$$

where

$$\frac{\partial l_i(\vartheta)}{\partial \beta} = A_i \mathcal{S}_{i1}(Y_i|X_i;\beta) + (1-A_i)E_{X_{i1}|Y_i,X_{i2}}[\mathcal{S}_{i1}(Y_i|X_i;\beta);\vartheta_1],$$

$$\frac{\partial l_i(\vartheta)}{\partial \gamma_1} = A_i \mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1) + (1-A_i)E_{X_{i1}|Y_i,X_{i2}}[\mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1);\vartheta_1],$$

and $\partial l_i(\vartheta)/\partial \gamma_2 = \mathcal{S}_{i3}(X_{i2};\gamma_2)$, respectively, where $\mathcal{S}_{i1}(Y_i|X_i;\beta) = \partial \log f(Y_i|X_i;\beta)/\partial\beta$, $\mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1) = \partial \log g_1(X_{i1}|X_{i2};\gamma_1)/\partial\gamma_1$, and $\mathcal{S}_{i3}(X_{i2};\gamma_2) = \partial \log g_2(X_{i2};\gamma_2)/\partial\gamma_2$. Note that the maximum likelihood approach requires modelling both the response $f(Y|X;\beta)$ and the nuisance distribution $g_1(X_1|X_2;\gamma_1)$ (Lawless et al., 1999); when X_1 is discrete the expectations above involve summation instead of integration. Note that the estimate of γ_2 is not needed for estimation of ϑ_1, but it has a role in approximating the asymptotic variance in the optimal design so we include it here.

We now consider the analysis at the conclusion of the adaptive two-phase sampling procedure where $B = 1$ indicates an individual, unselected in phase IIA, is selected in

phase IIB. Upon selecting M_B individuals from the $N_B = n - M_A$ eligible individuals following phase IIA, we have a phase IIB sub-sample. The score vector for ϑ_1 becomes

$$S_B(Y, X_1|X_2; \vartheta_1) = \sum_{i=1}^{n} S_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1) = \sum_{i=1}^{n} (S'_{i1B}(\vartheta_1), S'_{i2B}(\vartheta_1))',$$

where

$$\begin{aligned} S_{i1B}(\vartheta_1) &= R_i \mathcal{S}_{i1}(Y_i|X_i; \beta) + (1 - R_i) E_{X_{i1}|Y_i, X_{i2}}[\mathcal{S}_{i1}(Y_i|X_i; \beta); \vartheta_1] \\ S_{i2B}(\vartheta_1) &= R_i \mathcal{S}_{i2}(X_{i1}|X_{i2}; \gamma_1) + (1 - R_i) E_{X_{i1}|Y_i, X_{i2}}[\mathcal{S}_{i2}(X_{i1}|X_{i2}; \gamma_1); \vartheta_1] \ , \end{aligned} \quad (3.3)$$

where $R_i = A_i + (1 - A_i)B_i$, indicating that individual i is selected in either phase IIA or phase IIB, $i = 1, \ldots, n$. Under regularity conditions, the final estimate of ϑ at the end of the study, denoted by $\hat{\vartheta}$, obtained as the solution to the score equation $S_B(Y, X_1|X_2; \vartheta_1) = 0$, is asymptotically normal with

$$\sqrt{n}(\hat{\vartheta}_1 - \vartheta_1) \xrightarrow{D} N(0, E[S_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1) S'_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1)]^{-1}), \text{ as } n \to \infty, \quad (3.4)$$

where the asymptotic covariance matrix $E[S_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1) S'_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1)]$ involves expectations with respect to A, B, Y, X_1, and X_2. The joint expectations can be calculated as

$$E_{X_2}\{E_{X_1|X_2}\{E_{Y|X_1,X_2}\{E_{A,B|Y,X_1,X_2}[S_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1) S'_{iB}(Y_i, X_{i1}|X_{i2}; \vartheta_1)]\}\}\} \ ,$$

in which the expectation of X_1 given X_2 can be carried out by numerical integration when X_1 is continuous. The optimal $\hat{\psi}_B$ is the one that minimizes (3.2). After phase II selection is complete, analyses involve the score equation $S_B(Y, X_1|X_2; \vartheta_1) = 0$ for parameter estimates of interest. For inferential purposes, the expectations in the asymptotic covariance are replaced by empirical averages.

Inverse Probability Weighted Estimating Equations

Phase IIA parameter estimates can also be obtained by solving the inverse probability weighted estimating equations (IPWEE)

$$\sum_{i=1}^{n} U_{i1A}(Y_i|X_i; \beta, \psi_A) = \sum_{i=1}^{n} \frac{A_i}{\pi_{iA}} D'_{i1} \Sigma_{i1}^{-1} (Y_i - \mu_i) = 0$$

with $\mu_i = E(Y_i|X_i;\beta)$, $D_{i1} = \partial\mu_i/\partial\beta$, and $\Sigma_{i1} = \text{var}(Y_i|X_i;\beta)$. As the nuisance distribution $g_1(X_1|X_2;\gamma_1)$ is not modelled here, the nuisance parameter γ_1 is separately estimated as $\tilde{\gamma}_1$, the solution to the weighted score equations

$$\sum_{i=1}^{n} \frac{A_i}{\pi_{iA}} \mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1) = 0 \; ;$$

$\tilde{\gamma}_2$ solves $\sum_{i=1}^{n} \mathcal{S}_{i3}(X_{i2};\gamma_2) = 0$. After selection of the phase IIB sub-sample, we consider the IPWEE which marginalizes over the selection process

$$\sum_{i=1}^{n} U_{i1B}(Y_i|X_i;\beta,\psi^*) = \sum_{i=1}^{n} \frac{R_i}{\pi_i^*} D_{i1}' \Sigma_{i1}^{-1}(Y_i - \mu_i) = 0 \; , \tag{3.5}$$

together with score equation for the selection model parameters

$$\sum_{i=1}^{n} U_{i2B}(A_i, B_i;\psi^*) = \sum_{i=1}^{n} \frac{\partial\pi_i^*}{\partial\psi^*} \frac{1}{\pi_i^*(1-\pi_i^*)}(R_i - \pi_i^*) = 0 \; , \tag{3.6}$$

where we recall $R_i = A_i + (1-A_i)B_i$, and $\pi_i^* = \pi_{iA} + (1-\pi_{iA})\pi_{iB}$ is indexed by ψ^*. Thus, $U_{i1B}(Y_i|X_i;\beta,\psi^*)$ and $U_{i2B}(A_i,B_i;\psi^*)$ are functions of both ψ_A and ψ_B but with inputs ψ_A determined by the conventional design adopted for phase IIA. Under regularity conditions, the solution to (3.5) satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \Gamma_B^{-1}(I_B - H_B\Omega_B H_B')(\Gamma_B^{-1})'), \text{ as } n \to \infty, \tag{3.7}$$

where $\Gamma_B = E\left[-\partial U_{i1B}(Y_i|X_i;\beta,\psi^*)/\partial\beta\right]$, $I_B = E\left[U_{i1B}(Y_i|X_i;\beta,\psi^*)U_{i1B}'(Y_i|X_i;\beta,\psi^*)\right]$, $H_B = E[-\partial U_{i1B}(Y_i|X_i;\beta,\psi^*)/\partial\psi^*]$ and $\Omega_B = E\left[U_{i2B}(A_i,B_i;\psi^*)U_{i2B}'(A_i,B_i;\psi^*)\right]$, (Robins et al., 1995). The approximately optimal ψ_B then minimizes (3.2) as well.

The analysis after phase II involves solving the IPWEE (3.5) for parameter estimates without distinguishing ψ_A from ψ_B, but estimate the overall selection model π^* from the entire phase II sub-sample instead. Robins et al. (1994) and Lawless et al. (1999) explain this seemingly paradoxical gain in efficiency from estimating the selection probabilities.

The approach is in the spirit of McIssac and Cook (2015) who optimized the phase IIB selection model to minimize the variance of the IPW estimators based on the phase IIB sub-sample alone. In contrast, we consider the optimal phase IIB selection model in terms

of the overall estimator based on individuals who may have been chosen in phase IIA or IIB - thus we consider the potential effect of the sampling scheme in phase IIA. In McIssac and Cook (2015), the corresponding IPWEEs for the response and selection models in the design stage are

$$\sum_{i=1}^{n} U_{i1B}(Y_i|X_i;\beta,\psi) = \sum_{i=1}^{n} \frac{(1-A_i)B_i}{(1-\pi_{iA})\pi_{iB}} D'_{i1}\Sigma_{i1}^{-1}(Y_i-\mu_i) = 0,$$

and

$$\sum_{i=1}^{n} U_{i2B}(A_i,B_i;\psi) = \sum_{i=1}^{n} \frac{1-A_i}{1-\pi_{iA}} \frac{\partial \pi_{iB}}{\partial \psi} \frac{1}{\pi_{iB}(1-\pi_{iB})}(B_i-\pi_{iB}) = 0\ ,$$

respectively.

Conditional Likelihood

Scott and Wild (2011) advocated use of conditional likelihood in two-phase designs as they do not require modelling the nuisance distribution $g_1(x_1|X_2;\gamma_1)$. The conditional probability has the form

$$f(Y|X,A=1;\beta,\psi_A) = \frac{\pi_A(Y,X_2;\psi_A)f(Y|X;\beta)}{\sum_y \pi_A(y,X_2;\psi_A)f(y|X;\beta)}\ ,$$

and the score equations

$$\sum_{i=1}^{n} S^c_{iA}(Y_i|X_i;\beta,\psi_A) = \sum_{i=1}^{n} A_i \frac{\partial}{\partial\beta}\log f(Y_i|X_i,A_i=1;\beta,\psi_A) = 0\ ,$$

yield the phase IIA estimate $\tilde{\beta}$.

To construct the optimal phase IIB sub-sample in this framework, we consider the score equation from both phase IIA and IIB sub-samples,

$$\sum_{i=1}^{n} S^c_{iB}(Y_i|X_i;\beta,\psi) = \sum_{i=1}^{n} R_i \frac{\partial}{\partial\beta}\log f(Y|X,R_i=1;\beta,\psi) = 0\ , \tag{3.8}$$

where

$$f(Y|X,R=1;\beta,\psi) = \frac{[\pi_A(Y,X_2;\psi_A)+(1-\pi_A(Y,X_2;\psi_A))\pi_B(Y,X_2;\psi_B)]f(Y|X;\beta)}{\sum_y[\pi_A(y,X_2;\psi_A)+(1-\pi_A(y,X_2;\psi_A))\pi_B(y,X_2;\psi_B)]f(y|X;\beta)}.$$

The optimal ψ_B is the one that minimizes

$$E[S^c_{iB}(Y_i|X_i;\beta,\psi)S^{c'}_{iB}(Y_i|X_i;\beta,\psi);\tilde{\vartheta},\psi]^{-1}_{\beta_1,\beta_1} + \lambda\left[E(B|A=0;\tilde{\vartheta},\psi) - \frac{E(M_B)}{E(N_B)}\right]. \quad (3.9)$$

Note that although the score contribution is only a function of β and ψ, the joint expectations with respect to A, B, Y, X_1, and X_2 requires the nuisance distribution $g_1(x_1|X_2;\gamma_1)$. Finally, with both phase IIA and IIB sub-samples, analysis and inference after phase II are based on the score equation (3.8), with empirical averages employed whenever needed.

3.2.4 Robusteness and the Semiparametric Pseudo-Score

Chatterjee et al. (2003) proposed use of a semiparametric pseudo-score estimating function which exploits Bayes' rule to obtain a smoother and consistent nonparametric estimate of the nuisance distribution $g_1(X_1|X_2;\gamma_1)$ to improve the efficiency. As this approach does not require modelling the nuisance distribution $g_1(X_1|X_2;\gamma_1)$, robustness is achieved when X_1 is continuous.

Using Bayes' rule, for individuals selected in phase IIA, we have

$$g_1(x_1|X_2;\gamma_1) = \frac{dP(X_1 \le x_1|X_2, A=1)P(A=1|X_2)}{P(A=1|X_1=x_1,X_2)}\,. \quad (3.10)$$

Note that $P(X_1 \le x_1|X_2, A=1)$ can be replaced by an empirical estimate

$$\hat{G}_1(x_1|x_2, A=1) = \frac{\sum_{i=1}^n I(X_{i1} \le x_1, X_{i2}=x_2, A_i=1)}{\sum_{i=1}^n I(X_{i2}=x_2, A_i=1)}\,,$$

and the denominator of (3.10) is straightforward when Y is discrete, given by

$$P(A=1|X_1=x_1,X_2;\beta,\psi_A) = \sum_y \pi_A(y,X_2;\psi_A)f(y|X_1=x_1,X_2;\beta)\,.$$

Let $\tilde{g}_{1A}(X_1|X_2;\beta,\psi_A)$ denote the resultant estimate of the nuisance distribution $g_1(X_1|X_2;\gamma_1)$.

We can then write the score contribution in phase IIA as a function of β and ψ_A,

$$\begin{aligned}
S_A^p(Y|X;\beta,\psi_A) &= \sum_{i=1}^{n} A_i \frac{\partial}{\partial\beta} \log f(Y_i|X_i;\beta) g_1(X_{i1}|X_{i2};\gamma_1) \\
&\quad + (1-A_i)\frac{\partial}{\partial\beta} \log \int f(Y_i|x_1,X_{i2};\beta) g_1(x_1|X_{i2};\gamma_1) dx_1 \\
&= \sum_{i=1}^{n} A_i \mathcal{S}_{i1}(Y_i|X_i;\beta) \\
&\quad + (1-A_i)\frac{\int \mathcal{S}_{i1}(Y_i|x_1,X_{i2};\beta) f(Y_i|x_1,X_{i2};\beta)\tilde{g}_{1A}(x_1|X_{i2};\beta,\psi_A)dx_1}{\int f(Y_i|x_1,X_{i2};\beta)\tilde{g}_{1A}(x_1|X_{i2};\beta,\psi_A)dx_1} \\
&= \sum_{i=1}^{n} A_i \mathcal{S}_{i1}(Y_i|X_{i1},X_{i2};\beta) \\
&\quad + (1-A_i)\frac{\int \mathcal{S}_{i1}(Y_i|x_1,X_{i2};\beta) h(Y_i,x_1,X_{i2};\beta,\psi_A) dP(X_{i1}\le x_1|X_{i2},A_i=1)}{\int h(Y_i,x_1,X_{i2};\beta,\psi_A) dP(X_{i1}\le x_1|X_{i2},A_i=1)} ,
\end{aligned}$$

where

$$h(Y,X;\beta,\psi_A) = \frac{f(Y|X;\beta)}{P(A=1|X;\beta,\psi_A)} ,$$

and $P(X_1 \le x_1|X_2, A=1)$ is estimated empirically by $\hat{G}(x_1|X_2, A=1)$. Solving the score equations $S_A^p(Y|X;\beta,\psi_A) = 0$ gives the phase IIA estimates. Considering the phase IIA and phase IIB sub-samples, we write the nuisance distribution $g_1(x_1|X_2;\gamma_1)$ as

$$g_1(x_1|X_2;\gamma_1) = \frac{dP(X_1 \le x_1|X_2, R=1)P(R=1|X_2)}{P(R=1|X_1=x_1,X_2)} , \tag{3.11}$$

where $P(X_1 \le x_1|X_2, R=1)$ is estimated empirically as

$$\hat{G}_1(x_1|x_2,R=1) = \frac{\sum_{i=1}^{n} I(X_{i1}\le x_1, X_{i2}=x_2, R_i=1)}{\sum_{i=1}^{n} I(X_{i2}=x_2,R_i=1)} , \tag{3.12}$$

and the denominator as

$$\sum_y [\pi_A(y,X_2;\psi_A) + (1-\pi_A(y,X_2;\psi_A))\pi_B(y,X_2;\psi_B)] f(y|X_1=x_1,X_2;\beta) .$$

The score contribution is then a function of β and ψ^*

$$
\begin{aligned}
S_B^p(Y|X;\beta,\psi^*) &= \sum_{i=1}^{n} S_{iB}^p(Y_i|X_i;\beta,\psi^*) \qquad (3.13)\\
&= \sum_{i=1}^{n} R_i \frac{\partial}{\partial\beta} \log f(Y_i|X_{i1},X_{i2};\beta) g_1(X_{i1}|X_{i2};\gamma_1)\\
&+ (1-R_i)\frac{\partial}{\partial\beta}\log\int f(Y_i|x_1,X_{i2};\beta)g_1(x_1|X_{i2};\gamma_1)dx_1\\
&= \sum_{i=1}^{n} R_i \mathcal{S}_{i1}(Y_i|X_i;\beta)\\
&+ (1-R_i)\frac{\int \mathcal{S}_{i1}(Y_i|x_1,X_{i2};\beta)f(Y_i|x_1,X_{i2};\beta)\tilde{g}_{1B}(x_1|X_{i2};\beta,\psi^*)dx_1}{\int f(Y_i|x_1,X_{i2};\beta)\tilde{g}_{1B}(x_1|X_{i2};\beta,\psi^*)dx_1}\\
&= \sum_{i=1}^{n} R_i \mathcal{S}_{i1}(Y_i|X_i;\beta) + (1-R_i)\\
&\frac{\int \mathcal{S}_{i1}(Y_i|x_1,X_{i2};\beta)h(Y_i,x_1,X_{i2};\beta,\psi^*)dP(X_{i1}\le x_1|X_{i2},R_i=1)}{\int h(Y_i,x_1,X_{i2};\beta,\psi^*)dP(X_{i1}\le x_1|X_{i2},R_i=1)},
\end{aligned}
$$

with

$$
h(Y,X;\beta,\psi^*) = \frac{f(Y|X;\beta)}{P(R=1|X;\beta,\psi^*)}\ .
$$

Under regularity conditions (Chatterjee et al., 2003), the solution to the score equation $S_B^p(Y|X;\beta,\psi^*)=0$ asymptotically follows

$$
\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{D} N(0,(\mathcal{J}+\mathcal{C})^{-1}(\mathcal{J}+\Sigma)(\mathcal{J}'+\mathcal{C}')^{-1}),\ \text{as } n\to\infty, \qquad (3.14)
$$

where

$$
\mathcal{J} = -\frac{\partial}{\partial\beta}E[S_{iB}^p(Y_i|X_i;\beta,\psi^*)] = E[S_{iB}^p(Y_i|X_i;\beta,\psi^*)S_{iB}^{p'}(Y_i|X_i;\beta,\psi^*)]
$$

is the expected Fisher information matrix for the true likelihood,

$$
\mathcal{C} = E\left\{(1-\pi^*)\text{cov}\left[\mathcal{S}_1(Y|X;\beta),\frac{\partial}{\partial\beta}\log q(X;\beta,\psi^*)|Y,X_2\right]\right\}
$$

with the conditional selection probability $q(X;\beta,\psi^*)=P(R=1|X;\beta,\psi^*)$, and

$$
\Sigma = \text{var}[a_1(R,X)] + \mathcal{C} + \mathcal{C}' - \Psi\mathcal{J}_\psi^{-1}\Psi'
$$

with

$$a_1(R, X) = RE\left\{\frac{1-\pi^*}{q(X;\beta,\psi^*)}\{\mathcal{S}_1(Y|X;\beta) - E[\mathcal{S}_1(Y|X;\beta)|Y, X_2]\}|X\right\},$$

which adjusts the nonparametric estimate of $P(X_1 \leq x_1|X_2, R = 1)$ and is non-zero only for the selected individuals,

$$\mathcal{J}_\psi = E\left[\left(\frac{\partial\pi^*}{\partial\psi^*}\right)^2 \frac{1}{\pi^*(1-\pi^*)}\right],$$

and

$$\Psi = -E\left\{(1-\pi^*)\text{cov}\left[\mathcal{S}_1(Y|X;\beta), \frac{\partial}{\partial\psi^*}\log q(X;\beta,\psi^*)|Y, X_2\right]\right\}.$$

The term $\Psi\mathcal{J}_\psi^{-1}\Psi'$ adjusts for the estimation of π^*.

With the phase IIA estimates $(\tilde{\beta}', \tilde{\gamma}_2')'$, we aim to minimize (3.2) for a particular entry of the asymptotic covariance matrix in (3.14); each term in (3.14) involves an expectation $E(Z)$ and variance $E(ZZ') - E(Z)E(Z)'$ with respect to A, B, Y, X_1, and X_2 where Z represents an arbitrary vector-valued function of these random variables. Computation of the joint expectations requires numerical integration, but one can discretize X_1 in the phase IIA sample by computing empirical quantiles, binning the values and assigning the mean value to observations within bins. The continuous distribution $g_1(x_1|x_2;\gamma_1)$ then reduces to a discrete distribution $\hat{g}_1(x_1|x_2)$ with a finite number of points of support, enabling approximation of (3.10) by using the means at these points. Alternatively, $\tilde{g}_{1A}(x_1|x_2;\beta,\psi_A)$ can be estimated from the phase IIA sub-sample by collapsing to a piecewise constant function $\hat{g}_1(x_1|x_2)$ with a finite number of levels. Either choice of $\hat{g}$ can be used to speed up optimization in (3.2). The analysis and inference after phase II follow from solving $S_B^p(Y|X;\beta,\psi^*) = 0$ for the estimates of the parameter of interest.

Note that the semiparametric pseudo-score approach can be adapted to both the IP-WEE and conditional likelihood approaches. Although these approaches do not involve $g_1(x_1|X_2;\gamma_1)$ in their analysis, the semiparametric estimate helps in the design stage. Using the discretized version of the semiparametric estimator $\tilde{g}_{1A}(X_1|X_2;\beta,\psi_A)$, $\tilde{\gamma}_1$ is no longer needed in (3.2) and (3.9). Analysis and inference after phase II remain unchanged.

3.3 Empirical Studies

3.3.1 Design of Simulation Studies

The finite sample performance of estimators from the designs is investigated and compared in a variety of settings to demonstrate the efficiency and robustness. We consider a simulation study on $nsim = 1000$ generated phase I samples of size $n = 5000$. Such a phase I sample size represents scenarios of modern large cohort studies involving thousands of individuals from which a validation sub-sample may be chosen. For each individual, we consider a Bernoulli response Y whose response model is indexed by parameter $\beta = (\beta_0, \beta_1, \beta_2)'$,

$$E(Y|X_1, X_2; \beta) = \mu = \text{expit}(\beta_0 + \beta_1 X_1 + \beta_2 X_2), \tag{3.15}$$

where $\text{expit}(u) = e^u/(1 + e^u)$, and the parameter of interest is β_1. The Bernoulli auxiliary covariate X_2 satisfies

$$E(X_2; \gamma_2) = \text{expit}(\gamma_2). \tag{3.16}$$

When X_1 is Bernoulli, we set

$$E(X_1|X_2; \gamma_1) = \text{expit}(\gamma_{10} + \gamma_{11} X_2) \; ; \tag{3.17}$$

when X_1 is continuous, we suppose

$$X_1 = \gamma_{10} + \gamma_{11} X_2 + \epsilon \; , \tag{3.18}$$

where $\epsilon \sim N(0, \gamma_{12}^2)$. The parameter values are specified such that the marginal expectations of the response and covariates are all 0.20. When X_1 is binary, parameters are set to $(\beta_0, \beta_1, \beta_2, \gamma_{10}, \gamma_{11}, \gamma_2) = (-1.707, 0.916, 0.405, -1.753, 1.386, -1.386)$, and we consider the phase II sub-sample of expected size $E(M) = 500$ or $E(M) = 2000$ chosen via Bernoulli sampling. When X_1 is continuous, the parameter values are $(\beta_0, \beta_1, \beta_2, \gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_2) = (-2.121, 0.916, 0.405, 0, 1, \sqrt{2}, -1.386)$, and we consider the phase II sub-sample of expected size $E(M) = 500$ or $E(M) = 1000$ chosen via Bernoulli sampling. The binary response Y and auxiliary covariate X_2 lead to four strata based on the phase I sample, and the selection probabilities for phase IIA and IIB can be expressed as

$$\text{logit}\pi_A(Y, X_2; \psi_A) = \psi_{A0} + \psi_{A1} Y + \psi_{A2} X_2 + \psi_{A3} Y X_2,$$

and

$$\text{logit}\pi_B(Y, X_2; \psi_B) = \psi_{B0} + \psi_{B1}Y + \psi_{B2}X_2 + \psi_{B3}YX_2,$$

indexed by $\psi_A = (\psi_{A0}, \psi_{A1}, \psi_{A2}, \psi_{A3})'$ and $\psi_B = (\psi_{B0}, \psi_{B1}, \psi_{B2}, \psi_{B3})'$, respectively. In each set-up, the phase IIA sub-sample of expected size $E(M_A)$ employs simple random sampling (SRS) or balanced sampling (BS) for parameter estimation. The optimal phase IIB sub-sample has expected size $E(M_B) = E(M) - E(M_A)$, and its construction involves optimizing ψ_B. The optimization is done by using the `R` function `nlminb`. Derivatives required in (3.2) are computed using the function `fdHess` in `R` package `nmle`. Combining phase IIA and phase IIB sub-samples should give an consistent estimate of the parameter of interest $\hat{\beta}_1$. We assess robustness of inferences when X_1 is continuous by generating ϵ in (3.18) as $\epsilon \sim t(4)$ while retaining the mean and variance of $X_1|X_2$ as before. Here while the error term arises from a t-distribution, the design and analysis are based on the assumption that $X_1|X_2$ follows a normal distribution.

3.3.2 Empirical Findings from Simulation Studies

Empirical Findings from Simulation Studies with Binary X_1

When X_1 is binary, we compare the efficiencies of adaptive two-phase designs based on the maximum likelihood, IPWEE, and conditional likelihood. For each design, SRS or BS is employed in phase IIA with $E(M_A) = 0.25E(M)$ or $E(M_A) = 0.5E(M)$. Our suggestion is to find the optimal phase IIB selection model taking into account both the phase IIA and phase IIB selection processes, but we also compare this approach to designs that involve finding the phase IIB selection model to minimize the variance of the phase IIB estimator; within the IPWEE framework, this is the method proposed in McIssac and Cook (2015). Such design based on the maximum likelihood and conditional likelihood approaches have the scores in the form of (3.3) and (3.8), respectively, but with $A + (1 - A)B$ replaced as $(1 - A)B$.

The average standard errors (ASE) match the empirical standard errors (ESE) for the adaptive two-phase designs conducted within all frameworks of analysis and the empirical coverage probabilities (ECP) are compatible with the nominal 95% levels; see Table 3.1.

The advantages of adaptive two-phase designs over the non-adaptive standard SRS or BS designs are substantial when the phase II sub-sample size is small compared to the phase I sample size. This is not surprising as standard non-adaptive designs are not able to exploit the population well enough when the phase II sub-sample size is small. The likelihood approaches are comparable in terms of efficiency in all set-ups, and they are more efficient than the IPWEE approach. The implementation of the conditional likelihood is easier and in general leads to faster results due to the simplicity of the asymptotic covariance matrix. Within the framework of an IPWEE analysis, more efficient estimates are obtained if the phase IIB selection model is chosen based on the entire phase II sub-sample (see the "Full" columns) compared to the phase IIB estimator alone, with this gain increasing as the size of the phase IIA sub-sample increases. Smaller differences in efficiency are seen in likelihood analyses but despite this, the selection probabilities are quite different between the current approach and the approach of McIssac and Cook (2015); the optimal phase IIB selection probabilities $\hat{\pi}_B$ of the four strata, $(Y, X_2) = \{(0,0), (1,0), (0,1), (1,1)\}$, averaged over $nsim = 1000$ simulated datasets, are given in Table 3.2.

We define the asymptotic relative efficiency (ARE) of the adaptive two-phase designs under different phase IIA designs as

$$ARE = \frac{\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \vartheta, \psi_A, M_A]}{\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \vartheta, \psi, M]},$$

where the denominator represents a non-adaptive phase II design. Figure 3.2 plots the asymptotic standard error of $\hat{\beta}_1$ from the adaptive two-phase designs for likelihood and IPWEE analyses when simple random sampling (red) or balanced sampling (blue) are employed in phase IIA. Non-adaptive phase II SRS and BS designs are displayed as the upper bounds. The lower bound refers to the asymptotically-optimal phase II design which requires the true, but in practice, unknown parameters. As the size of the phase IIA sub-sample increases, the adaptive designs eventually approach the corresponding non-adaptive phase II designs. While true parameters were used to generate the curves, the ESEs from simulations with $E(M_A) = 0.1E(M)$ and $E(M_A) = 0.75E(M)$ agree with the plots. Still, we comment on the trade-off between the potential efficiency gain from a large phase IIB sub-sample and the loss in the precision of parameter estimates from a small phase IIA sub-sample. The best choice of M_A depends on the phase I sample size, phase II sub-sample

size, and the unknown parameters. Our simulation studies suggest exploiting a large phase IIB sub-sample yet avoid setting M_A too small to provide reasonable preliminary parameter estimates.

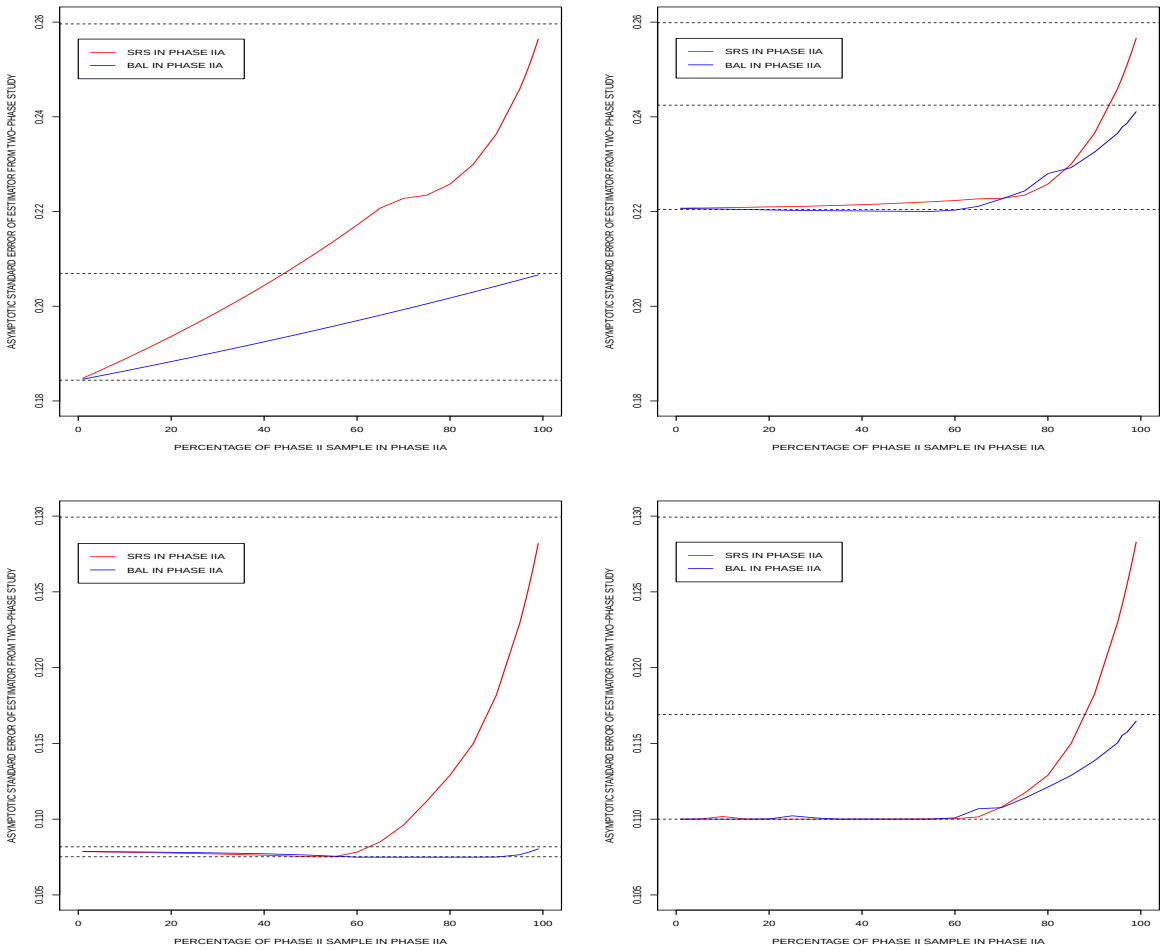

Figure 3.2: Plots of asymptotic standard error of $\hat{\beta}_1$ from likelihood (left) and IPWEE (right) adaptive two-phase designs. Red and blue curves indicate that SRS and BS are employed in phase IIA, respectively. The horizontal lines represent the asymptotic standard errors from the standard non-adaptive and asymptotically-optimal phase II designs. Phase I sample size $n = 5000$. Phase II sub-sample size $E(M) = 500$ in the first row and $E(M) = 2000$ in the second row. Parameter of interest $\beta_1 = 0.916$.

Table 3.1: Average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of estimators from maximum likelihood (ML), conditional likelihood (CML), and IPWEE (IPW) adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$ or $0.50E(M)$. Non-adaptive SRS and BS designs are included as the "100% IIA" columns. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$ (top half) or $E(M) = 2000$ (bottom half), and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

						Proposed Adaptive Designs					
			100% IIA			50% IIA			25% IIA		
Analysis	IIA	Opt[1]	ASE	ESE	ECP%	ASE	ESE	ECP%	ASE	ESE	ECP%
					$E(M) = 500$						
ML	SRS		0.264	0.261	95.5	0.212	0.211	94.8	0.198	0.202	94.1
	BS		0.208	0.204	95.7	0.196	0.194	96.3	0.191	0.192	95.4
CML	SRS		0.264	0.261	95.5	0.212	0.213	95.1	0.198	0.202	94.4
	BS		0.208	0.204	95.7	0.196	0.194	96.4	0.191	0.192	95.0
IPW	SRS	MC	0.264	0.261	95.5	0.231	0.239	93.6	0.226	0.229	95.9
		Full				0.224	0.229	94.2	0.223	0.229	94.6
	BS	MC	0.245	0.245	94.7	0.228	0.225	95.7	0.225	0.229	94.5
		Full				0.223	0.223	94.4	0.223	0.226	94.7
					$E(M) = 2000$						
ML	SRS		0.130	0.129	94.5	0.108	0.108	95.0	0.108	0.108	94.7
	BS		0.108	0.106	95.3	0.108	0.108	94.6	0.108	0.107	94.5
CML	SRS		0.130	0.129	94.5	0.108	0.108	95.3	0.108	0.108	94.5
	BS		0.108	0.106	95.3	0.108	0.108	94.8	0.108	0.106	95.0
IPW	SRS	MC	0.130	0.129	94.3	0.115	0.117	95.4	0.111	0.116	94.2
		Full				0.110	0.111	94.9	0.110	0.114	94.4
	BS	MC	0.117	0.114	96.2	0.112	0.111	95.6	0.111	0.113	94.8
		Full				0.110	0.110	95.0	0.111	0.114	94.7

[1] MC refers to approximation to phase IIB selection model based on phase IIB estimators only as in McIssac and Cook (2015). Full refers to that based on both phase IIA and IIB estimators.

Table 3.2: Optimal phase IIB selection probabilities for maximum likelihood (ML), conditional likelihood (CML), and IPWEE (IPW) estimators under adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$ or $0.50E(M)$. The "Full" columns refer to the designs optimizing the phase IIB sub-sample based on both phase IIA and IIB, while the "MC" columns refer to those based on phase IIB only. The expected strata sizes of $(Y, X_2) = \{(0,0), (1,0), (0,1), (1,1)\}$ are 3293, 708, 707, and 292, respectively. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$ (top half) or $E(M) = 2000$ (bottom half), and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

Phase IIA		(Y, X_2) Sampling Probability by Optimization Methods							
		MC				Full			
% Sampling*	Analysis	(0,0)	(1,0)	(0,1)	(1,1)	(0,0)	(1,0)	(0,1)	(1,1)
		$E(M) = 500$							
25 SRS	ML	0.001	0.001	0.275	0.651	0.001	0.001	0.268	0.666
	CML	0.001	0.001	0.271	0.659	0.001	0.001	0.264	0.676
	IPW	0.005	0.155	0.070	0.208	0.041	0.184	0.062	0.259
25 BS	ML	0.001	0.001	0.280	0.711	0.001	0.001	0.279	0.712
	CML	0.001	0.001	0.278	0.710	0.001	0.001	0.278	0.711
	IPW	0.050	0.161	0.069	0.214	0.055	0.175	0.051	0.163
50 SRS	ML	0.001	0.001	0.190	0.441	0.001	0.001	0.174	0.478
	CML	0.001	0.001	0.190	0.443	0.001	0.001	0.174	0.482
	IPW	0.034	0.109	0.046	0.142	0.190	0.162	0.029	0.225
50 BS	ML	0.001	0.001	0.194	0.543	0.001	0.001	0.197	0.535
	CML	0.001	0.001	0.194	0.543	0.001	0.001	0.196	0.536
	IPW	0.032	0.113	0.051	0.169	0.045	0.141	0.009	0.029
		$E(M) = 2000$							
25 SRS	ML	0.141	0.551	0.751	0.958	0.118	0.732	0.668	0.976
	CML	0.140	0.552	0.753	0.958	0.117	0.734	0.669	0.977
	IPW	0.204	0.725	0.331	0.837	0.166	0.885	0.335	0.937
25 BS	ML	0.182	0.833	0.478	0.954	0.174	0.712	0.639	0.954
	CML	0.180	0.834	0.477	0.954	0.173	0.713	0.640	0.955
	IPW	0.196	0.847	0.385	0.952	0.204	0.935	0.247	0.949
50 SRS	ML	0.069	0.276	0.778	0.948	0.010	0.680	0.625	0.997
	CML	0.069	0.276	0.779	0.948	0.010	0.682	0.625	0.997
	IPW	0.160	0.515	0.241	0.640	0.062	0.831	0.247	0.961
50 BS	ML	0.160	0.893	0.141	0.877	0.141	0.613	0.548	0.921
	CML	0.160	0.893	0.141	0.878	0.140	0.614	0.548	0.922
	IPW	0.134	0.848	0.349	0.952	0.189	0.797	0.074	0.576

* Percentage of the phase II sub-sample chosen from and the sampling scheme employed in phase IIA.

Empirical Findings from Simulation Studies with Continuous X_1

As expected, the results thus far have shown that likelihood approaches are more efficient than the methods based on IPWEE. This efficiency gain comes from modelling the nuisance covariate distribution $g_1(X_1|X_2;\gamma_1)$ which is at risk of misspecification. To investigate the impact of misspecification we consider next the case where ϵ in (3.18) follows a t-distribution with degree of freedom 4 but normality is assumed in both the analysis and the design.

We consider the phase II sub-sample size of expected size $E(M) = 500$ or $E(M) = 1000$, with the phase IIA sub-sample of expected size $E(M_A) = 0.25E(M)$. Tables 3.3 and 3.4 show the results from adaptive two-phase designs based on the maximum likelihood, the semiparametric pseudo-score method, the conditional likelihood, the conditional likelihood with the discretized semiparametric estimation applied in the design stage, IPWEE, and IPWEE with the discretized semiparametric estimation applied in the design stage; cases with SRS or BS employed in phase IIA are considered. The former table displays results of the setting with expected phase II sub-sample size $E(M) = 500$, and the latter table displays those of the setting with $E(M) = 1000$. While all designs are more efficient than the case of binary X_1, the standard balanced design has a close performance to the adaptive two-phase designs in the likelihood analysis frameworks. When the nuisance distribution is misspecified, the maximum likelihood approach yields biased estimators and poor coverage probabilities. This is not surprising since the maximum likelihood framework is susceptible to model misspecification, although modelling the nuisance covariates distributions can improve efficiency with likelihood analyses. The semiparametric pseudo-score approach, on the other hand, gives some improvements in terms of robustness, though the performance is unstable, particularly when the phase II sub-sample size is small. As the conditional likelihood and IPWEE approaches do not require the nuisance distribution $g_1(X_1|X_2;\gamma_1)$ in analysis, they are robust compared to the maximum likelihood. Despite this, adapting the semiparametric approach in the design drastically speeds up the computation and is beneficial when $g_1(X_1|X_2;\gamma_1)$ is misspecified.

Table 3.3: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of maximum likelihood (ML), conditional likelihood (CML), IPWEE (IPW), and semiparametric pseudo score (Semi-ML) estimators following adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$. Semiparametric conditional likelihood (Semi-CML) and semiparametric IPWEE (Semi-IPW) refer to combining the pseudo score estimation in the design stage with the analysis frameworks. Non-adaptive SRS and BS designs are included as the bottom "100%" rows. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$, and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

		Model For $X_1\|X_2$							
Phase IIA		Normal				Student (d.f. 4)			
% Sampling*	Analysis	EBias	ASE	ESE	ECP%	EBias	ASE	ESE	ECP%
25 SRS	ML	0.007	0.091	0.095	94.2	-0.218	0.091	0.084	32.2
	Semi-ML	0.002	0.074	0.100	84.1	0.002	0.079	0.104	85.9
	CML	0.008	0.095	0.097	94.5	0.021	0.106	0.102	96.3
	Semi-CML	0.009	0.095	0.097	94.7	0.019	0.104	0.100	96.5
	IPW	0.008	0.100	0.102	93.6	0.019	0.110	0.107	96.1
	Semi-IPW	0.008	0.098	0.100	93.3	0.019	0.108	0.107	94.6
25 BS	ML	0.006	0.091	0.093	95.0	-0.206	0.092	0.088	38.6
	Semi-ML	0.002	0.077	0.099	87.1	0.001	0.082	0.111	83.1
	CML	0.006	0.095	0.097	94.8	0.011	0.107	0.107	95.0
	Semi-CML	0.005	0.096	0.098	95.1	0.012	0.105	0.107	94.5
	IPW	0.006	0.099	0.099	95.6	0.013	0.109	0.108	94.6
	Semi-IPW	0.006	0.099	0.100	95.2	0.013	0.108	0.108	94.2
100 SRS	ML	0.014	0.116	0.120	95.4	0.005	0.125	0.123	94.7
	CML	0.017	0.116	0.120	95.3	0.012	0.126	0.124	95.2
	IPW	0.014	0.115	0.114	94.8	0.012	0.124	0.124	95.1
100 BS	ML	0.009	0.094	0.095	94.7	-0.128	0.098	0.097	72.5
	CML	0.009	0.096	0.097	94.4	0.010	0.106	0.108	93.8
	IPW	0.011	0.109	0.111	94.1	0.015	0.120	0.123	93.1

* Percentage of the phase II sub-sample chosen from and the sampling scheme employed in phase IIA.

Table 3.4: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) for maximum likelihood (ML), conditional likelihood (CML), IPWEE (IPW), and semiparametric pseudo score (Semi-ML) estimators under adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$. Semiparametric conditional likelihood (Semi-CML) and semiparametric IPWEE (Semi-IPW) refer to combining the pseudo score estimation in the design stage with the analysis frameworks. Non-adaptive SRS and BS designs are included as the bottom "100%" rows. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 1000$, and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

		Model For $X_1\|X_2$							
Phase IIA		Normal				Student (d.f. 4)			
% Sampling*	Analysis	EBias	ASE	ESE	ECP%	EBias	ASE	ESE	ECP%
25 SRS	ML	0.005	0.064	0.064	95.1	-0.225	0.064	0.062	7.10
	Semi-ML	0.001	0.070	0.070	94.5	-0.007	0.083	0.079	95.5
	CML	0.005	0.067	0.067	94.0	0.008	0.074	0.076	94.6
	Semi-CML	0.005	0.067	0.067	94.9	0.008	0.073	0.073	95.1
	IPW	0.008	0.070	0.070	94.8	0.009	0.078	0.079	94.8
	Semi-IPW	0.008	0.070	0.070	95.1	0.009	0.076	0.078	94.5
25 BS	ML	0.005	0.064	0.064	95.1	-0.212	0.065	0.068	11.6
	Semi-ML	−0.001	0.063	0.069	92.5	-0.006	0.074	0.079	93.4
	CML	0.006	0.067	0.068	94.2	0.011	0.076	0.077	94.9
	Semi-CML	0.006	0.068	0.069	93.7	0.009	0.073	0.075	95.2
	IPW	0.004	0.070	0.070	95.1	0.010	0.077	0.079	93.6
	Semi-IPW	0.004	0.070	0.071	95.0	0.010	0.076	0.078	93.8
100 SRS	ML	0.008	0.081	0.083	94.1	0.003	0.087	0.086	95.9
	CML	0.008	0.081	0.083	94.1	0.009	0.088	0.086	95.8
	IPW	0.008	0.080	0.083	94.2	0.008	0.087	0.086	94.9
100 BS	ML	0.002	0.066	0.064	95.7	-0.130	0.069	0.066	50.5
	CML	0.003	0.067	0.066	95.7	0.006	0.074	0.073	95.5
	IPW	0.004	0.078	0.077	95.0	0.008	0.087	0.091	93.3

* Percentage of the phase II sub-sample chosen from and the sampling scheme employed in phase IIA.

3.3.3 The Two-Phase Biomarker Study in Psoriatic Arthritis

Here we consider a motivating two-phase biomarker study in patients with psoriatic arthritis (PsA) where the goal is to most efficiently select individuals for the measurement of the biomarker matrix metalloproteinase 3 (MMP-3) in the University of Toronto Psoriatic Arthritis Clinic (UTPAC) to understand the association between MMP-3 levels and the disease progression while meeting budgetary constraints. A sample of 251 patients with MMP-3 measurements at a baseline assessment is available. Other relevant information from the data includes gender, erythrocyte sedimentation rate (ESR) at baseline, and clinical assessment results showing the number of clinical damaged joint counts at baseline and two years later. The severity of the clinical damaged joints is classified using the damage score which represents normal as grade 0, deformity as grade 1, ankylosis as grade 2, flail joint as grade 3, and surgery as grade 4.

We design a focused simulation study to investigate the performance of adaptive two-phase designs in the setting of the PsA program where we use pilot data help inform the parameter settings. We let X_1 denote the natural logarithm of the baseline MMP-3, and X_2 the dichotomized auxiliary covariate indicating whether there is an abnormal ESR measurement at the baseline assessment. The baseline ESR level is set to be 1 if it is greater than 20 for females or greater than 13 for males, and 0 otherwise according to the medical cut points. We let Y denote the binary response indicating whether the disease has progressed over two years of follow-up which is set to be 1 if there is an increase in the number of grade 1 or higher clinical damaged joints from the baseline assessment.

The response model and the auxiliary covariate model for X_2 have the form of (3.15) and (3.16), respectively. The conditional distribution $g_1(X_1|X_2;\gamma_1)$ satisfies

$$X_1 = I(X_2 = 0)\gamma_{10} + I(X_2 = 1)\gamma_{11} + \epsilon,$$

where

$$\epsilon \sim N(0, I(X_2 = 0)\gamma_{12}^2 + I(X_2 = 1)\gamma_{13}^2),$$

as the pilot data suggests that the log-normal distribution better characterizes the variation in MMP-3 levels. See Figure 3.3 for an illustration of the log-normality of the MMP-3 levels.

The configuration of the parameters β and γ_2 is obtained from fitting logistic regression models to the pilot data. The configuration of the parameter γ_1 is chosen to reflect the empirical distribution of the MMP-3 given the ESR levels in the pilot data. Specifically, $(\beta_0, \beta_1, \beta_2) = (-4.000, 0.320, -0.130)$, $(\gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_{13}) = (9.689, 10.084, 0.846, 1.151)$, and $\gamma_2 = -0.269$. We perform a simulation study with $nsim = 1000$, phase I sample size $n = 5000$, phase II sub-sample of expected size $E(M) = 500$, and phase IIA sub-sample of expected size $E(M_A) = 0.25E(M)$.

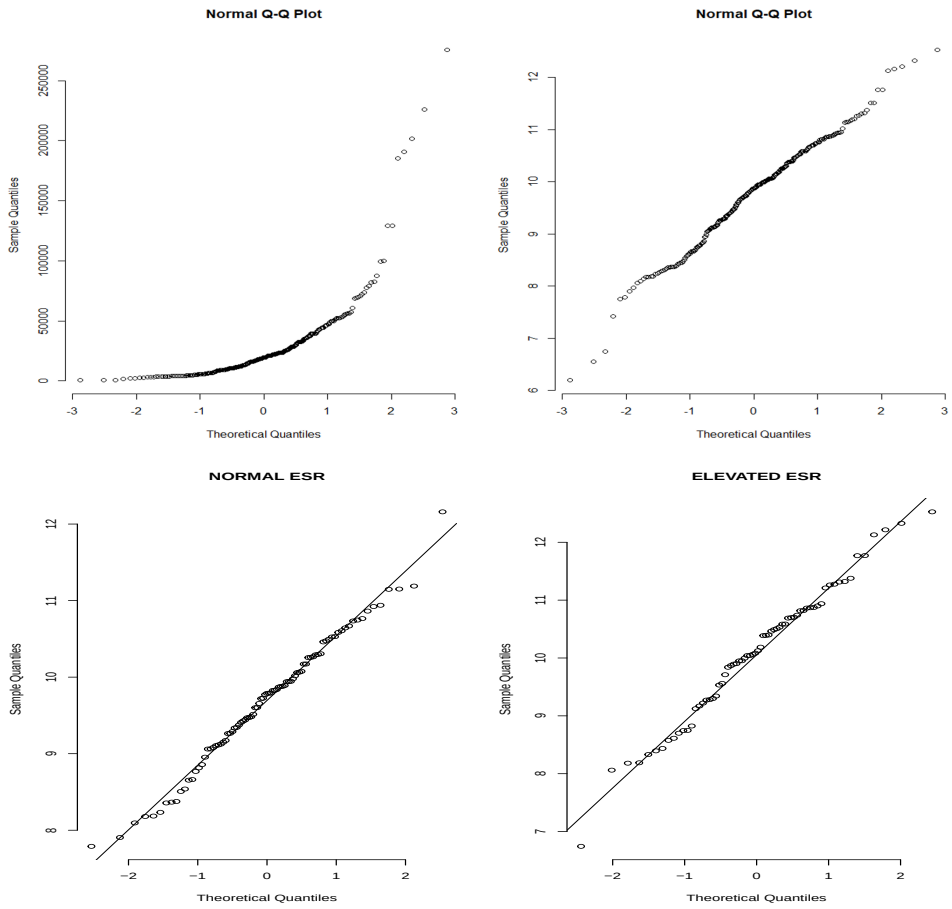

Figure 3.3: The first row displays normal QQ plots of MMP-3 levels (left) and their natural logarithms (right) from the pilot data. The second row displays QQ plots for the logged MMP-3 levels from the pilot data with normal (left) and elevated (right) ESR levels.

Table 3.5 shows the results of adaptive two-phase designs conducted via the conditional likelihood and IPWEE approaches with discretized semiparametric estimation of the nuisance distribution in the design stage. Traditional SRS or BS is employed in phase IIA. The proposed adaptive two-phase designs avoid the need to model the nuisance distribution $g_1(X_1|X_2;\gamma_1)$ and are more efficient than standard SRS or BS designs; this is particularly true when analyses are planned based on conditional likelihood. Hence, the proposed designs are well adapted to the PsA setting to inspect the relationship between the disease progression and biomarkers of interest with budget limitations.

Table 3.5: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of conditional likelihood (Semi-CML) and IPWEE (Semi-IPW) estimators from adaptive two-phase designs of the PsA study setting with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$. The analysis frameworks are combined with the semiparametric estimation of the nuisance distribution in the design. Non-adaptive SRS and BS designs are included as the "100% IIA" columns. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$, and $nsim = 1000$. The response is whether there is an increase in the number of grade 1 or higher damaged joints in two years of follow-up and the parameter of interest $\beta_1 = 0.320$.

						Proposed Adaptive Designs			
		100% IIA				25% IIA			
Analysis	IIA Sampling	EBias	ASE	ESE	ECP%	EBias	ASE	ESE	ECP%
Semi-CML	SRS	0.009	0.104	0.103	95.4	0.005	0.086	0.083	96.2
	BS	0.008	0.093	0.095	94.8	0.004	0.085	0.084	95.0
Semi-IPW	SRS	0.009	0.104	0.103	95.6	0.008	0.094	0.093	95.6
	BS	0.008	0.094	0.096	94.4	0.004	0.094	0.093	95.3

3.4 The Surrogate Value Problem

Here we consider the use of an adaptive two-phase design to address the surrogate variable problem in which $Y \perp X_2|X_1$. While we only adjust for the exposure variable X_1, the auxiliary covariate X_2 still helps determine the weights and sampling probabilities. Such designs can arise in cases where X_1 is a definitive test result and X_2 is an inexpensive inaccurate result which we only use to sample. In other words, the response model reduces to

$$E(Y|X_1;\beta) = \mu = \text{expit}(\beta_0 + \beta_1 X_1),$$

with the nuisance distributions $g_1(X_1|X_2;\gamma_1)$, $g_2(X_2;\gamma_2)$, and selection models $\pi(Y, X_2;\psi)$ unchanged. Hence, the approaches presented in Section 3.2 have the same form except that we remove parameter β_2 completely from the response model μ - we neither estimate β_2 from the phase IIA sub-sample, nor include β_2 in the objective function of the optimization. It is not surprising that the adaptive two-phase designs of a surrogate value problem have different efficiencies and optimal phase IIB selection probabilities $\hat{\pi}_B$ from those adjusting both X_1 and X_2. Table 3.6 reports the results of the adaptive two-phase designs of a surrogate value problem based on the maximum likelihood, conditional likelihood, and IPWEE approaches from simulation studies with the same set-up as in Section 3.3.2 for binary X_1. Adaptive two-phase designs improve efficiency of estimation over standard non-adaptive designs, and the efficiency gain is more appreciable when the phase II sub-sample size is small. When X_1 is binary, designs based on maximum likelihood are the most efficient. Moreover, it is also more efficient to optimize the phase IIB sub-sample based on both phase IIA and phase IIB individuals (the "Full" columns) in the IPWEE analysis framework. Table 3.7 displays results from simulation studies with the same set-up as in Section 3.3.2 for continuous X_1. When X_1 is continuous, the adaptive two-phase designs of a surrogate value problem based on the conditional likelihood and IPWEE analysis frameworks combining the semiparametric estimation in the design are both efficient and robust to the misspecification of the nuisance distribution, though the improvement from a standard non-adaptive BS is not evident in the conditional likelihood analysis framework.

Table 3.6: Average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of estimators from maximum likelihood (ML), conditional likelihood (CML), and IPWEE (IPW) adaptive two-phase designs of a surrogate value problem with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$ or $0.5E(M)$. Non-adaptive SRS and BS designs are included as the "100% IIA" columns. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$ (top half) or $E(M) = 2000$ (bottom half), and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

						Proposed Adaptive Designs					
			100% IIA			50% IIA			25% IIA		
Analysis	IIA	Opt[1]	ASE	ESE	ECP%	ASE	ESE	ECP%	ASE	ESE	ECP%
					$E(M) = 500$						
ML	SRS		0.204	0.205	95.8	0.181	0.182	95.5	0.171	0.175	94.8
	BS		0.173	0.167	95.5	0.163	0.162	95.5	0.162	0.156	96.4
CML	SRS		0.254	0.255	95.1	0.204	0.209	94.4	0.192	0.200	93.4
	BS		0.197	0.192	95.0	0.190	0.188	95.7	0.187	0.181	96.7
IPW	SRS	MC	0.254	0.255	95.1	0.222	0.225	94.2	0.217	0.215	95.9
		Full				0.215	0.213	95.5	0.214	0.212	95.3
	BS	MC	0.236	0.232	95.5	0.219	0.215	95.0	0.216	0.216	96.1
		Full				0.214	0.210	95.6	0.214	0.212	95.9
					$E(M) = 2000$						
ML	SRS		0.119	0.118	95.0	0.102	0.103	95.6	0.102	0.102	95.0
	BS		0.102	0.102	94.7	0.102	0.102	95.6	0.102	0.100	94.7
CML	SRS		0.126	0.125	95.5	0.104	0.104	95.4	0.104	0.104	95.3
	BS		0.104	0.104	95.0	0.104	0.104	96.1	0.104	0.103	95.4
IPW	SRS	MC	0.126	0.122	96.0	0.110	0.113	95.4	0.107	0.110	94.4
		Full				0.106	0.107	95.7	0.106	0.109	94.4
	BS	MC	0.112	0.114	94.4	0.110	0.110	95.0	0.107	0.106	95.1
		Full				0.107	0.107	94.8	0.106	0.104	95.2

[1] MC refers to approximation to phase IIB selection model based on phase IIB estimators only as in McIssac and Cook (2015). Full refers to that based on both phase IIA and IIB estimators.

Table 3.7: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) for maximum likelihood (ML), conditional likelihood (CML), and IPWEE (IPW) estimators under adaptive two-phase designs of a surrogate value problem with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$. The analysis frameworks are combined with the semiparametric estimation of the nuisance distribution in the design. Non-adaptive SRS and BS designs are included as the bottom "100%" rows. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 500$ (top half) or $E(M) = 1000$ (bottom half), and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

		Model For $X_1\|X_2$							
Phase IIA		Normal				Student (d.f. 4)			
% Sampling*	Analysis	EBias	ASE	ESE	ECP%	EBias	ASE	ESE	ECP%
		$E(M) = 500$							
25 SRS	Semi-CML	0.006	0.092	0.097	93.5	0.020	0.100	0.101	94.6
	Semi-IPW	0.008	0.094	0.098	93.9	0.017	0.102	0.099	94.5
25 BS	Semi-CML	0.006	0.091	0.088	96.1	0.013	0.098	0.095	96.2
	Semi-IPW	0.007	0.095	0.093	96.5	0.015	0.102	0.102	95.4
100 SRS	CML	0.017	0.111	0.115	94.9	0.012	0.119	0.120	94.5
	IPW	0.017	0.110	0.112	94.8	0.010	0.118	0.117	95.1
100 BS	CML	0.003	0.089	0.088	95.6	0.012	0.097	0.093	95.8
	IPW	0.007	0.102	0.100	95.7	0.019	0.111	0.108	95.4
		$E(M) = 1000$							
25 SRS	Semi-CML	0.004	0.061	0.061	94.4	0.007	0.065	0.065	94.6
	Semi-IPW	0.006	0.066	0.066	94.5	0.007	0.071	0.073	94.5
25 BS	Semi-CML	0.003	0.064	0.065	94.5	0.007	0.068	0.067	95.3
	Semi-IPW	0.005	0.067	0.068	95.2	0.009	0.072	0.073	95.2
100 SRS	CML	0.007	0.078	0.080	93.9	0.008	0.083	0.082	95.1
	IPW	0.007	0.077	0.078	94.0	0.007	0.083	0.081	95.9
100 BS	CML	0.002	0.063	0.061	95.9	0.007	0.068	0.066	95.6
	IPW	0.004	0.073	0.070	96.5	0.011	0.079	0.080	95.6

* Percentage of the phase II sub-sample chosen from and the sampling scheme employed in phase IIA.

3.5 Discussion

In this chapter, we have investigated the performance of adaptive response-dependent two-phase sampling schemes for discrete or continuous expensive exposure variables. With the phase II sub-sample size fixed by budgetary constraints, we recommend the most efficient adaptive two-phase designs based on maximum likelihood - concerns regarding misspecification of the nuisance covariate distribution are less serious when the expensive exposure variable is discrete but the models warrant checking. When the exposure is continuous, conditional likelihood analysis incorporating semiparametric estimation of the nuisance distribution at the design stage is recommended to enhance efficiency while maintaining robustness. The designs can be adapted to a surrogate value problem in which we only adjust the expensive exposure variable in the response model. The investigations in Appendix 3A do not suggest that partitioning the phase II selection procedure into multiple (> 2) stages leads to meaningful improvements in performance so we do not advocate it for the kinds of settings we considered.

The adaptive approach to the selection of the phase II sample holds promise for more complex settings where information may be lacking about key parameters, and where standard phase II selection models may be unappealing due to this complexity. Future work would include designing adaptive two-phase sampling schemes involving longitudinal responses which may require knowledge of the dependence structure of the response. If exposure variables are time-varying, then the most appropriate approach to stratification is unclear, however, for optimal phase II sampling a model would be required to describe the joint evolution of the covariate of interest and the response. An interim phase II sample could be very valuable in such a case. Alternative approaches for two-phase sampling warranting development include calibration (Rivera-Rodriguez et al., 2019) and adopting informative priors of the parameters before phase IIA sampling (Chen and Lumley, 2020). The proposed adaptive designs approximate the optimal model-based design to approach the maximaly efficient estimators. Further development of design-based approaches considered by Chen and Lumley (2020) is an important area of future research.

Chapter 4

Secondary Analysis and Sequential Design of Two-Phase Studies

4.1 Introduction

4.1.1 Background and Literature Review

Biomarker studies involving two-phase response-dependent sampling have great appeal when large cohort studies entail the collection and storage of biospecimens and interest lies in assessing the role of a biomarker in disease onset or progression. Sampling strategies are developed to select biospecimens, which will be assayed to observe the biomarker of interest and obtain an efficient estimate of its effect while respecting budgetary constraints. Given that modern consortiums of cohort studies create platforms for the conduct of multiple biomarker studies, it is often favourable to perform analyses leveraging the exposure variables collected from earlier studies to address new but possibly related scientific questions. The idea of reusing collected exposure data involves using it for the secondary purpose of an upcoming study. So far, most of the work on secondary use of available data, referred to as secondary analysis in literature, concentrates on case-control studies. Lin and Zeng (2009) developed likelihood methods to analyse secondary phenotype data in case-control studies,

which extended to a retrospective likelihood framework in Ghosh et al. (2013). Tchetgen (2014) examined the re-parameterization of the conditional model for the secondary response given the case-control status and regression covariates. Assuming linear models for the secondary response, Pan et al. (2018a) proposed an estimated likelihood approach for secondary analysis based on time-to-event data in case-cohort studies via joint modelling. See Schifano (2019) for a comprehensive review on this topic. As for cohort studies, Saarela et al. (2012) discussed a conditional likelihood approach for secondary analysis with a time-to-event response of interest. Pan et al. (2018b) used estimating equations to perform secondary analysis in response-dependent sampling designs.

We consider sequential two-phase designs conducted on the same platform. We develop a response-dependent two-phase design that examines the association between a previously studied biomarker and a new response. Such designs can address scenarios in which the interest is to study the role of certain biomarkers in the progression of one disease, with some of the serum samples in the registry assayed from earlier studies that investigated the effect of the biomarkers on the progression of other diseases. When researchers run out of budget in an upcoming study, no additional serum samples will be assayed after the earlier studies, and the problem reduces to a secondary analysis in the context of two-phase designs. Otherwise, we expect the responses and auxiliary covariates of the previous studies to provide extensive information to maximize statistical efficiency in the upcoming study subject to a new set of budgetary constraints. The exposure variables measured from earlier studies help approximate the optimal sampling scheme in the upcoming study in the spirit of an adaptive two-phase design (McIssac and Cook, 2015). In either case, we adopt joint response models and perform analysis via maximum likelihood, conditional likelihood, and inverse probability-weighted estimating equations. In Section 4.1.2, we give an overview of the framework for sequential two-phase designs of our interest.

4.1.2 Notation and Framework for Sequential Two-Phase Designs

In this section, we introduce necessary notation and establish a framework of sequential two-phase studies. We consider two sequential two-phase studies without loss of generality.

Our objective is to study the effect of a reused exposure from one study on the response of interest in the second study. Consider a two-phase study, which we refer to as Study 1 from now on, that is planned to model $\mu_1(X, Z) = E(Y_1|X, Z)$, where Y_1 denotes a response, X the expensive exposure variable of interest, and Z the vector of discrete auxiliary covariates. Suppose the response and auxiliary data are available from a cohort study or registry comprised of n independent individuals, which form the phase I sample of Study 1, denoted by $\mathcal{D}_{11} = \{Y_{i1}, Z_i : i \in \mathcal{R}\}$, where $\mathcal{R} = \{1, \ldots, n\}$. We note that the methods we discuss can be readily adapted to deal with the surrogate variable problem in which interest lies in modelling $E(Y_1|X)$. Budgetary constraints, and the need to preserve biospecimens, preclude the measurement of X on all individuals, and we suppose only some fraction of individuals in the phase I sample can have their biospecimens assayed in Study 1.

Much of the work on phase II sub-sampling of two-phase designs involves stratification based on the phase I sample and the use of stratum-specific selection probabilities. Let $C_1(Y_1, Z)$ be a coarse mapping from (Y_1, Z') onto K_1 strata, labeled as $\{s_{1j} : j = 1, \ldots, K_1\}$, so that $C_1(Y_{i1}, Z_i)$ records the stratum to which individual i belongs in Study 1. The mapping $C_1(\cdot)$ determines the nature of the stratification. We let $R_{i1} = 1$ if individual i is selected in phase II of Study 1 so that X_i is observed, where R_{i1} is realized based on selection model $\pi_{i1} = \pi_1(Y_{i1}, Z_i) = P(R_{i1} = 1|Y_{i1}, Z_i; \psi_1)$. The selection model may correspond to simple random sampling, proportional, balanced (Breslow and Cain, 1988) or some other specified sampling schemes. In this framework, the covariate X is therefore missing at random (MAR) (Little and Rubin, 2002). Let $\mathcal{R}_1 = \{i \in \mathcal{R} : R_{i1} = 1\}$ indicate the set of individuals chosen for the phase II sub-sample of Study 1 with $M_1 = |\mathcal{R}_1|$ the size of the phase II sub-sample, and let $\mathcal{R}_1^c = \{i \in \mathcal{R} : R_{i1} = 0\}$ consist of those individuals who are not selected. The data upon completion of phase II is therefore $\mathcal{D}_{12} = \{Y_{i1}, X_i, Z_i : i \in \mathcal{R}_1\} \cup \{Y_{i1}, Z_i : i \in \mathcal{R}_1^c\}$.

Following the conduct of Study 1, we consider a subsequent two-phase study, referred to as Study 2, which is planned to model $\mu_2(X, Z) = E(Y_2|X, Z)$, where Y_2 is a new response and it is completely observed in $\mathcal{R}$. The phase I data of Study 2 is $\mathcal{D}_{21} = \{Y_{i1}, Y_{i2}, X_i, Z_i : i \in \mathcal{R}_1\} \cup \{Y_{i1}, Y_{i2}, Z_i : i \in \mathcal{R}_1^c\}$, and we suppose that the selection model π_1 used for Study 1 is known at the planning stage of Study 2. Suppose that the new budget allows for the selection of some fraction of individuals in $\mathcal{R}_1^c$ to create the Study 2

phase II sub-sample. Note that following Study 1, only $n - M_1$ individuals remain eligible, as the M_1 individuals already have their expensive covariates collected. With a coarse mapping $C_2(Y_1, Y_2, Z)$ from (Y_1, Y_2, Z') onto K_2 strata $\{s_{2j} : j = 1, \ldots, K_2\}$, we let $R_{i2} = 1$ if individual i is selected in phase II of Study 2. The selection indicator R_{i2} is realized based on selection model $\pi_{i2} = \pi_2(Y_{i1}, Y_{i2}, Z_i) = P(R_{i2} = 1|Y_{i1}, Y_{i2}, Z_i, R_{i1} = 0; \psi_2)$. Let $\mathcal{R}_2 = \{i \in \mathcal{R} : R_{i1} = 0, R_{i2} = 1\}$ and $M_2 = |\mathcal{R}_2|$. The data acquisition steps for a sequence of two-phase studies then follow, of which an illustration is given in Figure 4.1. We here focus on the second study in such a sequence. Upon completion of Study 2, individuals in $\mathcal{R}_1 \cup \mathcal{R}_2$ have X available with M_1 of them chosen by π_1 and M_2 of them chosen by π_2. For maximum statistical efficiency, an optimal Study 2 aims to select individuals from $\mathcal{R}_1^c$ to minimize the variability of the estimator of the parameter of interest from the combined data $\mathcal{D}_{22} = \{Y_{i1}, Y_{i2}, X_i, Z_i : i \in \mathcal{R}_1 \cup \mathcal{R}_2\} \cup \{Y_{i1}, Y_{i2}, Z_i : i \in \mathcal{R} \setminus (\mathcal{R}_1 \cup \mathcal{R}_2)\}$ subject to budgetary constraints. The mapping $C_2(\cdot)$ defining the stratification in Study 2 facilitates the formation of a class of selection models within which such optimal designs will be considered. This framework can be easily generalized to settings involving more than two studies by treating previous studies combined as Study 1 and a new study as Study 2.

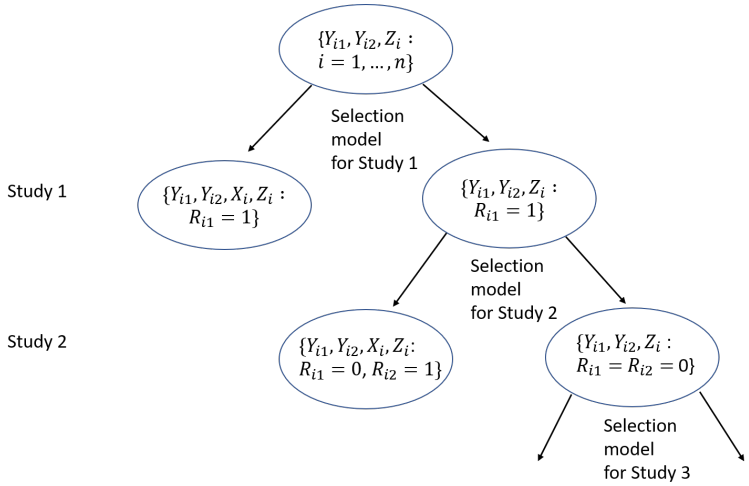

Figure 4.1: A schematic representing a series of two-phase studies based on a common platform cohort study; following completion of Study 2 individuals in $\mathcal{R}_1 \cup \mathcal{R}_2$ have available data with M_1 chosen by selection model $\pi_1(Y_1, Z)$ and M_2 chosen by $\pi_2(Y_1, Y_2, Z)$.

The remainder of the chapter is structured as follows. In Section 4.2, coping with an exhausted budget, we introduce a joint response model followed by outlining the secondary analysis in the context of two-phase designs. In Section 4.3, subject to new budgetary constraints, we propose the most efficient sequential two-phase designs and investigate the finite sample performance of estimators from the designs in various settings. In Section 4.4, we look into robustness issues and the performances of the methods when the joint response model is misspecified. In Section 4.5, we apply our methods to a real-life dataset aiming to investigate the association between a Human Leukocyte Antigen biomarker and the development of joint damage in a psoriatic arthritis research program. Section 4.6 concludes with some general remarks.

4.2 A Framework for Secondary Analysis of Two-Phase Studies

In this section, we delineate the idea of reusing the exposure variables measured in Study 1 to analyse the new response of interest Y_2. This may be of interest when there is no budget available for the collection of additional data on X. The joint response model, introduced in Section 4.2.1, is adopted in the frameworks of analysis described in Section 4.2.2.

4.2.1 A Joint Response Model

Upon completion of Study 1, there are M_1 individuals whose exposure variables are available for secondary analyses of Y_2. However, since the selection model π_1 for Study 1 is based on $C_1(Y_1, Z)$, for standard analysis of the model for $Y_2|X, Z$ the exposure X may be missing not at random (MNAR). Appropriate secondary analyses strategies require characterization of the relationship between Y_1 and Y_2 so a joint model for $Y|X, Z$ is of interest, $Y = (Y_1, Y_2)'$. For illustration, we consider here the case where Y_1 and Y_2 are binary with

$$\mu_1(X, Z) = E(Y_1|X, Z) = \text{expit}(\alpha_0 + \alpha_1 X + \alpha_2' Z), \tag{4.1}$$

and

$$\mu_2(X, Z) = E(Y_2|X, Z) = \text{expit}(\beta_0 + \beta_1 X + \beta_2' Z), \tag{4.2}$$

respectively, where $\alpha = (\alpha_0, \alpha_1, \alpha_2')'$, $\beta = (\beta_0, \beta_1, \beta_2')'$, β_1 is the parameter of primary interest, and $\text{expit}(u) = \exp(u)/(1 + \exp(u))$. The conditional dependence between Y_1 and Y_2 can be characterized by the odds ratio

$$\phi(X, Z) = \frac{P(Y = (1, 1)'|X, Z)/P(Y = (0, 1)'|X, Z)}{P(Y = (1, 0)'|X, Z)/P(Y = (0, 0)'|X, Z)}, \tag{4.3}$$

and modeled via

$$\log \phi(X, Z) = \gamma_0 + \gamma_1 X + \gamma_2' Z + \gamma_3' XZ, \tag{4.4}$$

a generalization to the bivariate logistic model in Palmgren (1989). We further let $\gamma = (\gamma_0, \gamma_1, \gamma_2', \gamma_3')'$ and $\theta = (\alpha', \beta', \gamma')'$. The joint model for the responses and covariates is

$$P(Y, X, Z; \vartheta) = P(Y|X, Z; \theta) f(X|Z; \xi) f(Z; \zeta), \tag{4.5}$$

where $\vartheta = (\theta', \xi', \zeta')'$, and ξ and ζ are key parameters for shaping the design of the phase II sub-sampling scheme.

4.2.2 Secondary Analysis of Data from an Earlier Two-Phase Study

In this section, we discuss strategies to conduct secondary analysis of exposure data measured in Study 1 for the new response of interest including inverse probability-weighted estimating equations (IPWEE), maximum likelihood, and conditional maximum likelihood.

Inverse Probability Weighted Estimating Equations

Here we consider second-order inverse probability weighted (IPW2) estimating equations (Prentice, 1988; Zhao and Prentice, 1990) modelling both the mean and dependence parameters. This approach does not exploit information from the unselected individuals but

weights estimating equations based on the selection model used in Study 1. The estimating functions are therefore of the form

$$\bar{U}_1^J(\theta, \psi_1) = \sum_{i=1}^{n} \bar{U}_{i1}^J(\theta, \psi_1) = \sum_{i=1}^{n} \frac{R_{i1}}{\pi_{i1}(Y_{i1}, Z_i; \psi_1)} U_{i1}^J(\theta) \ , \tag{4.6}$$

where $U_{i1}^J(\theta) = D_i \Sigma_i^{-1} B_i$,

$$D_i = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \alpha} & 0 & \frac{\partial \eta_i}{\partial \alpha} \\ 0 & \frac{\partial \mu_{i2}}{\partial \beta} & \frac{\partial \eta_i}{\partial \beta} \\ 0 & 0 & \frac{\partial \eta_i}{\partial \gamma} \end{pmatrix} \ , \qquad \Sigma_i = \begin{pmatrix} \text{cov}(Y_{i1}, Y_{i2}|X_i, Z_i)^{-1} & 0 \\ 0 & \text{var}^{-1}(Y_{i1}Y_{i2}|X_i, Z_i) \end{pmatrix} \ , \tag{4.7}$$

and $B_i = (Y_{i1} - \mu_{i1}, Y_{i2} - \mu_{i2}, Y_{i1}Y_{i2} - \eta_i)'$ with

$$\begin{aligned} \eta(X, Z) &= E(Y_1 Y_2|X, Z) \\ &= \frac{a(X, Z) - \{a(X, Z)^2 - 4\phi(X, Z)[\phi(X, Z) - 1]\mu_1(X, Z)\mu_2(X, Z)\}^{0.5}}{2[\phi(X, Z) - 1]}, \end{aligned}$$

in which

$$a(X, Z) = 1 - [1 - \phi(X, Z)][\mu_1(X, Z) + \mu_2(X, Z)],$$

$\mu_{i1} = \mu_1(X_i, Z_i)$, $\mu_{i2} = \mu_2(X_i, Z_i)$, and $\eta_i = \eta(X_i, Z_i)$. We adopt a block diagonal working covariance matrix to avoid modeling higher-order dependence parameters and note that $\text{var}(Y_1 Y_2|X, Z) = \eta(X, Z) - \eta(X, Z)^2$ since the responses are binary. While the selection model is known at this point, efficiency gains are realized if the selection model is fitted by solving

$$U_1^R(\psi_1) = \sum_{i=1}^{n} \frac{\partial \pi_{i1}}{\partial \psi_1} \frac{1}{\pi_{i1}(1 - \pi_{i1})}(R_{i1} - \pi_{i1}) = 0$$

and using the estimated weights in (4.6); see remarks in Section 4.3.1. If we set $\partial \eta_i / \partial \alpha = \partial \eta_i / \partial \beta = 0$ in (4.7), no information about the marginal parameters is obtained from the estimating functions regarding dependence parameters which reduce the IPW2 to the weighted first-order estimating equations (IPW) and in turn strengthen the robustness of inference regarding β to misspecification of the dependence structure; see Section 4.4 for details. The IPWEEs are expected to be robust but inefficient compared to the following likelihood methods.

Maximum Likelihood

Since the selection process is non-informative (i.e., ϑ and ψ_1 are functionally independent), the observed data likelihood

$$L_1^J(\vartheta) = \prod_{i=1}^{n} [P(Y_i|X_i, Z_i; \theta) f(X_i|Z_i; \xi) f(Z_i; \zeta)]^{R_{i1}} P(Y_i, Z_i; \vartheta)^{1-R_{i1}}, \tag{4.8}$$

can be used for inference. The log-likelihood is

$$\begin{aligned} l_1^J(\vartheta) = \sum_{i=1}^{n} R_{i1} [\log P(Y_i|X_i, Z_i; \theta) + \log f(X_i|Z_i; \xi)] \\ + (1 - R_{i1}) \log P(Y_i|Z_i; \vartheta) + \log f(Z_i; \zeta). \end{aligned}$$

The maximum likelihood (ML) estimate is obtained by solving the score equations

$$\sum_{i=1}^{n} \begin{pmatrix} S_{i1\theta}^J(\vartheta_1) \\ S_{i1\xi}^J(\vartheta_1) \\ S_{i1\zeta}^J(\zeta) \end{pmatrix} = 0,$$

with

$$\begin{aligned} S_{i1\theta}^J(\vartheta_1) &= R_{i1} \mathcal{S}_{i\theta}(Y_i|X_i, Z_i) + (1 - R_{i1}) E\left[\mathcal{S}_{i\theta}(Y_i|X_i, Z_i)|Y_i, Z_i\right]; \\ S_{i1\xi}^J(\vartheta_1) &= R_{i1} \mathcal{S}_{i\xi}(X_i|Z_i) + (1 - R_{i1}) E\left[\mathcal{S}_{i\xi}(X_i|Z_i)|Y_i, Z_i\right]; \\ S_{i1\zeta}^J(\zeta) &= \partial \log f(Z_i)/\partial\zeta, \end{aligned}$$

where we let $\vartheta_1 = (\theta', \xi')'$, and $\mathcal{S}_{i\theta}(Y_i|X_i, Z_i) = \partial \log P(Y_i|X_i, Z_i)/\partial\theta$ and $\mathcal{S}_{i\xi}(X_i|Z_i) = \partial \log f(X_i|Z_i)/\partial\xi$ are the complete data score functions. Note that the full likelihood depends on both the response model and the nuisance covariate models. Lawless et al. (1999) pointed out that modelling the covariate distributions can improve efficiency with likelihood analyses, but this framework is susceptible to model misspecification.

Conditional Maximum Likelihood

Under conditional maximum likelihood (CML) we restrict attention to the phase II subsample of Study 1, $\mathcal{R}_1$, and solve the score equation

$$S_{1\theta}^{CJ}(\theta, \psi_1) = \sum_{i=1}^{n} R_{i1} \frac{\partial \log P(Y_i|X_i, Z_i, R_{i1} = 1)}{\partial \theta} = 0, \tag{4.9}$$

where

$$\begin{aligned} P(Y_i|X_i, Z_i, R_{i1} = 1) &= \frac{P(R_{i1} = 1|Y_i, X_i, Z_i)P(Y_i|X_i, Z_i)}{P(R_{i1} = 1|X_i, Z_i)} \\ &= \frac{\pi_1(Y_{i1}, Z_i)P(Y_i|X_i, Z_i)}{\sum_{y_1} \pi_1(y_1, Z_i)P(y_1|X_i, Z_i)}. \end{aligned}$$

Scott and Wild (2011) recommended this approach for its simplicity and the fact that it does not require estimation of nuisance parameters indexing covariate distributions. The avoidance of modelling the covariate distributions in the analysis brings some robustness, and conditioning on the minimal sufficient statistic for nuisance parameters renders little efficiency loss for parameters indexing the response models.

4.2.3 Empirical Studies of Approaches to Analysis of Secondary Responses

We now assess the performance of the proposed secondary analyses via simulation studies. The responses and covariates are taken to be scalar binary random variables. The covariates satisfy

$$E(Z) = \text{expit}(\zeta) \text{ and } E(X|Z) = \text{expit}(\xi_0 + \xi_1 Z). \tag{4.10}$$

The responses are modelled marginally as in (4.1) and (4.2) together with the association addressed as in (4.3). We specify $\xi_1 = \log 4$ to reflect a strong association between the exposure of interest and the auxiliary variable Z, $\alpha_1 = \beta_1 = \log 2.5$ to reflect strong effects of the exposure on the responses, and $\alpha_2 = \beta_2 = \log 1.5$ to reflect modest effects of the auxiliary covariate on the responses. By further specifying

$(\zeta, \xi_0, \alpha_0, \beta_0) = (-1.386, -1.753, -1.707, -1.707)$, we ensure that the marginal probabilities of both the covariates and responses are 0.2. Finally, regarding the dependence model we set $\gamma_1 = \gamma_2 = \gamma_3 = 1$ in (4.4), and the marginal odds ratio of the full cohort,

$$E[\phi(X, Z)] = \sum_{x,z} E(\phi(x, z)|X = x, Z = z)P(X = x, Z = z),$$

to be 2 or 4, reflecting moderate and strong associations between responses and leading to $\gamma_0 = -0.395$ and $\gamma_0 = 0.298$, respectively. The size of the phase I sample of Study 1 is taken as $n = 5000$, and the phase II sub-sample of expected size $E(M_1) = 250$ is chosen via Bernoulli sampling. For each parameter setting, $nsim = 1000$ simulations are carried out. For illustration, we assume that Study 1 employs balanced sampling (BS) such that an equal number of individuals are chosen from each of four strata defined by (Y_1, Z). This can be expressed through a selection model of the form

$$\text{logit}\pi_1(Y_1, Z; \psi_1) = \psi_{10} + \psi_{11}Y_1 + \psi_{12}Z + \psi_{13}Y_1Z$$

with $\psi_1 = (\psi_{10}, \psi_{11}, \psi_{12}, \psi_{13})'$ chosen accordingly.

The analysis methods described in Section 4.2.2 are based on joint response models, but marginal models may also be fitted. For the inverse weighting method in Section 4.2.2, we therefore also consider

$$U_1^M(\beta, \psi_1) = \sum_{i=1}^{n} \frac{R_{i1}}{\pi_{i1}(Y_{i1}, Z_i; \psi_1)} \frac{\partial \mu_{i2}}{\partial \beta} \text{var}(Y_{i2}|X_i, Z_i)^{-1}(Y_{i2} - \mu_{i2}) = 0, \tag{4.11}$$

which involves use of weights from Study 1 applied to a score function for $Y_2|X, Z$ marginally; as this is the correct weight function, the resulting estimator is expected to have small empirical bias and valid inference following use of a robust standard error. The likelihood simplifies to

$$L_1^M(\beta, \xi, \zeta) = \prod_{i=1}^{n} [P(Y_{i2}|X_i, Z_i; \beta)f(X_i|Z_i; \xi)f(Z_i; \zeta)]^{R_{i1}} P(Y_{i2}, Z_i; \beta, \xi, \zeta)^{1-R_{i1}};$$

but we note that here the exposure data are MNAR so inconsistent estimates are expected.

The simulation results reported in Table 4.1 are given separately for the joint and marginal analysis frameworks. "Joint" and "Marginal" stand for adopting the joint response model and the marginal model in analysis, respectively. The average standard

errors (ASE) are computed via established asymptotic theory for the corresponding analysis frameworks; see Section 4.3.1 for details. As expected the marginal analysis yields estimators with negligible finite sample biases for IPW but significant biases for ML. The empirical biases are small for all estimators resulting from joint analyses as these are valid approaches. Among these, ML yields the most efficient estimator, followed by CML and then the IPWEEs. The joint analyses based on IPW and IPW2 are slightly more efficient than the corresponding marginal method. Finally, we observe smaller empirical biases and standard errors when there is a strong association between Y_1 and Y_2; see results of $E[\phi(X, Z)] = 4$ compared to those of $E[\phi(X, Z)] = 2$. As the association between the response used for sampling (Y_1) and the response of interest (Y_2) gets stronger, the secondary analysis gets closer to an ordinary analysis in a standard two-phase study.

4.3 Efficient Sequential Two-Phase Designs

In this section, we consider the scenario in which the available data are to be used to help shape the design of a new study. We propose how to maximize statistical efficiency while meeting budgetary constraints in a second two-phase study given the exposure variables collected from an earlier study conducted on the same platform.

4.3.1 Design and Analysis of a Sequence of Two-Phase Studies

We aim to construct and evaluate selection models for individuals not selected for the phase II sub-sample of Study 1 (i.e., those in the set $\mathcal{R}_1^c$). An optimal selection model is expected to select individuals to maximize statistical efficiency within the corresponding analysis framework. For a given approach to analysis, the optimal selection model governs the phase II sub-sampling scheme of Study 2 that yields an estimator of β_1 with the minimum asymptotic variance among sub-samples with the same expected size. Such designs, however, require knowledge of unknown parameters ϑ indexing the joint model (4.5) for the responses and covariates. The data from Study 1 may be viewed as pilot data yielding preliminary estimate denoted as $\tilde{\vartheta}$ based on the approaches described in Section

Table 4.1: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the parameter estimates of interest from the IPW and IPW2 (Section 4.2.2), ML (Section 4.2.2), and CML (Section 4.2.2) secondary analyses following a two-phase study based on Y_1 (Study 1) employing balanced sampling with $n = 5000$ and $E(M_1) = 250$.

Response Model	$E[\phi(X, Z)]$	Analysis	EBias	ASE	ESE	ECP(%)
Joint	2	IPW	0.021	0.434	0.448	92.9
		IPW2	0.023	0.432	0.443	92.6
		ML	0.021	0.337	0.344	94.2
		CML	0.029	0.356	0.353	96.2
	4	IPW	0.008	0.423	0.436	93.2
		IPW2	0.004	0.427	0.428	92.8
		ML	0.008	0.329	0.340	93.2
		CML	0.013	0.347	0.349	94.3
Marginal	2	IPW	0.028	0.442	0.461	93.0
		ML	0.386	0.317	0.318	76.4
	4	IPW	0.013	0.428	0.454	92.5
		ML	0.357	0.309	0.320	79.1

4.2. This estimate can help approximate the optimal phase II selection model for Study 2 with an approximately optimal $\pi_2 = P(R_2 = 1|Y, Z, R_1 = 0; \psi_2)$ defined as the one that minimizes

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \tilde{\vartheta}, \psi_1, \psi_2] + \lambda \left[E(R_2|R_1 = 0; \tilde{\vartheta}, \psi_1, \psi_2) - \frac{E(M_2)}{n - E(M_1)} \right], \tag{4.12}$$

where λ denotes the Lagrange multiplier, ψ_1 known from Study 1, and the parameter estimates $\tilde{\vartheta}$ are substituted into the expressions. The objective function and the constraint are functions of ψ_2.

We next outline the procedure of how the optimal selection model of Study 2 can be conducted in different frameworks of analysis.

Inverse Probability Weighted Estimating Equations

Upon the completion of Study 2, the two subsequent phase II sub-samples will be combined as $\mathcal{D}_{22}$ for analysis. The IPW2 can be written as

$$\bar{U}_2^J(\theta, \bar{\psi}_2) = \sum_{i=1}^{n} \bar{U}_{i2}^J(\theta, \bar{\psi}_2) = \sum_{i=1}^{n} \frac{\bar{R}_{i2}}{\bar{\pi}_{i2}} U_{i2}^J(\theta) = 0, \tag{4.13}$$

where $U_{i2}^J(\theta) = D_i \Sigma_i^{-1} B_i$ with D_i, Σ_i, and B_i defined in Section 4.2.2. Here $\bar{R}_{i2} = R_{i1} + (1 - R_{i1})R_{i2}$ and $\bar{\pi}_{i2} = \pi_{i1} + (1 - \pi_{i1})\pi_{i2}$ is indexed by $\bar{\psi}_2$. Since π_1 is known in advance, we treat ψ_1 as fixed. Therefore, $\bar{\psi}_2$ as a function of both ψ_1 and ψ_2 is determined by ψ_2. The estimating equation for the net selection model is

$$U_2^R(\bar{\psi}_2) = \sum_{i=1}^{n} U_{i2}^R(\bar{\psi}_2) = \sum_{i=1}^{n} \frac{\partial \bar{\pi}_{i2}}{\partial \bar{\psi}_2} \frac{1}{\bar{\pi}_{i2}(1 - \bar{\pi}_{i2})} (\bar{R}_{i2} - \bar{\pi}_{i2}) = 0. \tag{4.14}$$

Under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \Gamma^{-1}(I - H\Omega H')(\Gamma^{-1})'),$$

as $n \to \infty$, where matrices $\Gamma = E[-\partial \bar{U}_{i2}^J(\theta, \bar{\psi}_2)/\partial\theta']$, $I = E\{\bar{U}_{i2}^J(\theta, \bar{\psi}_2)[\bar{U}_{i2}^J(\theta, \bar{\psi}_2)]'\}$, $\Omega = E\{U_{i2}^R(\bar{\psi}_2)[U_{i2}^R(\bar{\psi}_2)]'\}$, and $H = E[-\partial \bar{U}_{i2}^J(\theta, \bar{\psi}_2)/\partial\bar{\psi}_2']$. The approximately optimal $\bar{\psi}_2$ is the approximately optimal ψ_2 that minimizes (4.12). The joint expectation with respect to Y, X, Z, R_1, and R_2 can be evaluated as $E_{X,Z}\{E_{Y|X,Z}[E_{R_1,R_2|Y,X,Z}(\cdot)]\}$. With a preliminary estimate $\tilde{\theta}$ obtained from (4.6), the nuisance parameters ξ and ζ are still required in the design though they are not involved in the analysis. Their preliminary estimates can be obtained separately, with $\tilde{\xi}$ by weighted logistic regression and $\tilde{\zeta}$ by score equations $S_{i1\zeta}(\zeta) = \partial \log P(Z_i)/\partial\zeta$. Note that the reduction in the asymptotic variance by term $H\Omega H'$ represents the efficiency gain from estimating the known $\bar{\psi}_2$ upon the completion of Study 2. Such a seemingly paradoxical efficiency gain from estimating the known selection probabilities was denoted as "estimation better than the known" and explained in previous research (Lawless et al., 1999; Robins and Rotnitzky, 1995). Moreover, the corresponding IPW follows immediately by setting $\partial\eta_i/\partial\alpha = \partial\eta_i/\partial\beta = 0$ in D_i, with the hope to enhance robustness. The net selection model (4.14) remains unchanged, and the resulting asymptotic covariance matrix of estimators of θ still has the sandwich form.

Maximum Likelihood

Upon the completion of Study 2, the score equations become

$$S_2^J(\vartheta_1) = \sum_{i=1}^{n} S_{i2}^J(\vartheta_1) = \sum_{i=1}^{n} \begin{pmatrix} S_{i2\theta}^J(\vartheta_1) \\ S_{i2\xi}^J(\vartheta_1) \end{pmatrix} = 0, \tag{4.15}$$

where

$$S_{i2\theta}^J(\vartheta_1) = \bar{R}_{i2}\mathcal{S}_{i\theta}(Y_i|X_i, Z_i) + (1 - \bar{R}_{i2})E[\mathcal{S}_{i\theta}(Y_i|X_i, Z_i)|Y_i, Z_i],$$

and

$$S_{i2\xi}^J(\vartheta_1) = \bar{R}_{i2}\mathcal{S}_{i\xi}(X_i|Z_i) + (1 - \bar{R}_{i2})E[\mathcal{S}_{i\xi}(X_i|Z_i)|Y_i, Z_i].$$

Under regularity conditions, standard asymptotic theory gives

$$\sqrt{n}(\hat{\vartheta}_1 - \vartheta_1) \xrightarrow{D} N(0, E\{S_{i2}^J(\vartheta_1)[S_{i2}^J(\vartheta_1)]'\}^{-1}),$$

as $n \to \infty$. Calculation of the asymptotic variance requires knowledge of the joint model of the response and covariates (4.5) and parameters ϑ indexing the model. We obtain preliminary parameter estimates $\tilde{\vartheta}$ from likelihood (4.8) using the exposure variables measured in Study 1. With ψ_1 known from Study 1, the optimal ψ_2 is the one that minimizes (4.12).

Conditional Maximum Likelihood

Upon the completion of Study 2, the score equations become

$$S_2^{CJ}(\theta, \psi) = \sum_{i=1}^{n} \bar{R}_{i2} \frac{\partial \log P(Y_i|X_i, Z_i, \bar{R}_{i2} = 1)}{\partial \theta} = 0, \tag{4.16}$$

where

$$\begin{aligned} P(Y_i|X_i, Z_i, \bar{R}_{i2} = 1) &= \frac{P(\bar{R}_{i2} = 1|Y_i, Z_i, X_i)P(Y_i|X_i, Z_i)}{P(\bar{R}_{i2} = 1|X_i, Z_i)} \\ &= \frac{[\pi_1(Y_{i1}, Z_i) + \pi_2(Y_i, Z_i)(1 - \pi_1(Y_{i1}, Z_i))]P(Y_i|X_i, Z_i)}{\sum_{y_1,y_2}[\pi_1(y_1, Z_i) + \pi_2(y_1, y_2, Z_i)(1 - \pi_1(y_1, Z_i))]P(y_1, y_2|X_i, Z_i)}. \end{aligned}$$

Under regularity conditions, standard asymptotic theory holds and the optimal design refers to the ψ_2 that minimizes (4.12). The asymptotic covariance matrix is different from

that of the ML approach because of the distinct score equations. Similarly, preliminary parameter estimates $\tilde{\theta}$ can be obtained from (4.9), and $\tilde{\xi}$ and $\tilde{\zeta}$ are obtained from logistic regression and score equations, respectively.

4.3.2 Empirical Studies of Sequential Two-Phase Designs

The finite sample performance of the design is investigated in a variety of settings and compared to alternative routinely adopted approaches to demonstrate the efficiency gains via empirical studies of sizes $nsim = 1000$, $n = 5000$, and $E(M_1) = E(M_2) = 250$. Bernoulli sampling is used. We adopt the parameter configuration in Section 4.2.3. While we assume that Study 1 employs BS based on 4 strata defined by Y_1 and Z, our proposed sequential design with optimal efficiency for Study 2 is conducted based on 8 strata defined by Y_1, Y_2, and Z with selection model

$$\begin{aligned}\text{logit}\pi_2(Y, Z; \psi_2) = \psi_{20} + \psi_{21}Y_1 + \psi_{22}Y_2 + \psi_{23}Z \\ + \psi_{24}Y_1Z + \psi_{25}Y_2Z + \psi_{26}Y_1Y_2 + \psi_{27}Y_1Y_2Z,\end{aligned} \tag{4.17}$$

and is compared to the following designs.

A. Independent Design for Study 2. Here we consider the case where there is no communication regarding information of X acquired in Study 1, and hence, this information is not utilized in Study 2. We assume, however, that the biospecimen samples have been preserved and so individual samples chosen in Study 2 can be assayed. Thus, we select M_2 individuals using BS based on 4 strata defined by Y_2 and Z. This is for sure inefficient as it completely fails to utilize any information from Study 1. A standard two-phase design analysis involving $P(Y_2|X, Z)$ is performed via ML, CML, and IPW. For IPW, one could consider

$$U_2^M(\beta, \psi_2) = \sum_{i=1}^{n} \frac{R_{i2}}{\pi_{i2}(0, Y_{i2}, Z_i)} \frac{\partial \mu_{i2}}{\partial \beta} \text{var}(Y_{i2}|X_i, Z_i)^{-1}(Y_{i2} - \mu_{i2}) = 0. \tag{4.18}$$

As for the likelihood analyses, we have

$$L_2^M(\beta, \xi, \zeta) = \prod_{i=1}^{n} [P(Y_{i2}|X_i, Z_i; \beta) f(X_i|Z_i; \xi) f(Z_i; \zeta)]^{R_{i2}} P(Y_{i2}, Z_i; \beta, \xi, \zeta)^{1-R_{i2}} \tag{4.19}$$

for ML, and

$$S_2^{CM}(\theta, \psi_2) = \sum_{i=1}^{n} R_{i2} \frac{\partial \log P(Y_{i2}|X_i, Z_i, R_{i2} = 1)}{\partial \theta} = 0 \tag{4.20}$$

for CML. Note that R_2 and π_2 arise from standard BS in such designs.

B. Naive Marginal Analysis in Study 2. Although the exposure variables measured from Study 1 are included for analysis, Study 2 simply performs BS to select M_2 individuals from the $n - M_1$ individuals in $\mathcal{R}_1^c$ based on 4 strata defined by Y_2 and Z. Indeed, the combined phase II sub-sample has a larger size of $M_1 + M_2$, and Study 2 avoids repetitive selections. However, such designs ignore the correlation between the responses and perform the same analyses as in design **A** involving the marginal model $P(Y_2|X, Z)$. The IPW and likelihood analyses follow from (4.18) to (4.20), where R_2 and π_2 are replaced by $\bar{R}_2 = R_1 + (1 - R_1)R_2$ and $\bar{\pi}_2 = \pi_1 + (1 - \pi_1)\pi_2$, respectively, while here π_1 and π_2 arise from BS employed in Study 1 and Study 2, respectively.

C. Study 2 Design Based on Joint Response Model. While including the exposure variables measured from Study 1 for analysis, Study 2 adopts the joint response model and performs BS to select M_2 individuals from $\mathcal{R}_1^c$ based on 8 strata defined by Y_1, Y_2 and Z. Such selection and the joint analysis involving $P(Y|X, Z)$ allow for a valid aggregation of the $M_1 + M_2$ exposure variables from two sequential studies. See (4.13), (4.15), and (4.16) for the conduct of such designs in IPW and IPW2, ML, and CML, respectively. Nevertheless, the estimates may not be the most efficient ones as the selection model in (4.17) still arises from standard BS.

D. Approximate Optimal Designs. The proposed selection model in (4.17) whose ψ_2 optimizes (4.12) defines our optimal design which we label as $\mathbf{D_1}$. Another optimal design selects $M_1 + M_2$ individuals for Study 2 from $\mathcal{R}$ to achieve maximum efficiency when investigating $P(Y_2|X, Z)$. The IPW and likelihood analyses follow from (4.18) to (4.20), but with ψ_2 minimizing

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \vartheta, \psi_2] + \lambda \left[E(R_2; \vartheta, \psi_2) - \frac{E(M_1) + E(M_2)}{n} \right]. \tag{4.21}$$

This design, labelled as $\mathbf{D_2}$, is ideal but unrealistic because we assume that we know the information of unknown parameters of the population ϑ.

Table 4.2 display the results of designs **A - D** conducted via likelihood and inverse weighting methods. Designs **A** are inefficient as expected since they have smaller phase II sub-sample sizes. Still, they are more efficient than the results of the secondary analyses in Table 4.1, since the individuals are selected based on Y_2 and Z after all. Designs **B** that fail to incorporate information on Y_1 in design and analyses lead to biased estimates. Designs **C**, adopting the joint response model, give unbiased, though not necessarily very efficient, estimates. Our proposed optimal designs $\mathbf{D_1}$ are valid and more efficient. The empirical biases (EBias) are small for designs **C** and $\mathbf{D_1}$ conducted via all frameworks of analysis. The average standard errors (ASE) match the empirical standard errors (ESE), and the empirical coverage probabilities (ECP) are compatible with the nominal 95% levels. Moreover, our proposed optimal designs $\mathbf{D_1}$ are found to be close in terms of efficiency to the infeasible designs $\mathbf{D_2}$. For an illustration across the phase II sub-sample sizes of the sequential studies, Figure 4.2 plots the asymptotic standard error of $\hat{\beta}_1$ of our proposed optimal designs $\mathbf{D_1}$ against the proportion of the individuals in the combined phase II sub-sample that are selected from Study 1, i.e., $E(M_1)/E(M_1+M_2)$. True parameters are used to calculate the asymptotic standard errors when generating the plots. As $E(M_1)/E(M_1+M_2)$ increases, the efficiency of our proposed designs $\mathbf{D_1}$ approaches to that of employing BS in Study 1 to select M_1+M_2 individuals. On the other hand, as the proportion decreases, the efficiency attains that of the unrealistic designs $\mathbf{D_2}$ which optimally select M_1+M_2 individuals in Study 2.

Table 4.3 summarizes the sampling probabilities of the 8 strata defined by Y and Z, referred to as

$$(Y_1, Y_2, Z) = \{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,0,1),(0,1,1),(1,1,0),(1,1,1)\},$$

of designs $\mathbf{D_1}$ and $\mathbf{D_2}$ displayed in Table 4.2. As designs $\mathbf{D_2}$ focus on Y_2 to optimally select M_1+M_2 individuals in Study 2, they consider 4 strata of (Y_2, Z) and the sampling probabilities of strata $(1, Y_2, Z)$ are equal to those of strata $(0, Y_2, Z)$. Our proposed designs $\mathbf{D_1}$, on the other hand, involve BS with fixed $\pi_1(Y_1, Z; \psi_1)$ in Study 1 and an optimal $\pi_2(Y, Z; \psi_2)$ in Study 2. Sampling probabilities π_1, π_2, and $\bar{\pi}_2$ are displayed in the "Study 1", "Study 2", and "Net Study 2" rows, respectively. For likelihood approaches, designs $\mathbf{D_2}$ have extreme sampling probabilities. While it is not surprising for the "Net Study 2"

rows to differ from the $\mathbf{D_2}$ rows, our proposed designs $\mathbf{D_1}$ make efforts to save the mess created by the BS scheme. For example, the optimal designs $\mathbf{D_1}$ avoid selecting additional individuals from strata $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(1,1,0)$ in Study 2, given that designs $\mathbf{D_2}$ select very few individuals from those strata. As for the IPW approach, the sampling probabilities of designs $\mathbf{D_2}$ are not that extreme. The "Net Study 2" rows display results that roughly match the $\mathbf{D_2}$ rows.

In terms of the analyses, ML is the most efficient as it models the distributions of the exposure variables. However, we found that CML is computationally much faster due to its simpler score equations and is almost as efficient as ML. The IPW and IPW2 methods are not as efficient as the likelihood approaches. IPW2 does not show clear improvements from IPW. Nonetheless, such improvements are found to be substantial in a set-up where the responses have an exchangeable dependence structure. See Table 4A1 in Appendix 4A for the simulation results of such a setting in which we specify $\phi(X,Z) = 2$ and $\phi(X,Z) = 4$ for a moderate and a strong association between the responses, respectively. The theoretical asymptotic standard error plots and the sampling probabilities of strata (Y,Z) of the designs are also available in Appendix 4A as Figure 4A1 and Table 4A2, respectively, where similar results are found.

Table 4.2: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the estimator of parameter of interest from the combined phase II data using likelihood and inverse weighting methods. **A-D** refer to designs with different use of Study 1 data described in Section 4.3.2. For designs **A-D$_1$**, $E(M_1) = E(M_2) = 0.05n$. For designs **D$_2$**, $E(M_2) = 0.1n$. "BS" and "opt" stand for balanced sampling and optimal sampling, respectively. Moderate and strong associations between the responses are reflected by marginal odds ratios 2 and 4, respectively. $nsim = 1000$, $n = 5000$, and $\beta_1 = 0.916$.

	Stratification				Results			
Design	Study 1	Study 2	Response Model	Analysis	EBias	ASE	ESE	ECP(%)
			$E[\phi(X,Z)] = 2$					
A	-	BS (Y_2, Z)	Marginal	IPW	0.030	0.350	0.352	95.6
				ML	0.011	0.298	0.290	95.6
				CML	0.011	0.298	0.290	95.5
B	BS (Y_1, Z)	BS (Y_2, Z)	Marginal	IPW	0.046	0.241	0.245	95.1
				ML	0.078	0.207	0.207	93.4
				CML	0.078	0.207	0.207	93.4
C	BS (Y_1, Z)	BS (Y, Z)	Joint	IPW	-0.001	0.275	0.280	94.4
				IPW2	0.001	0.274	0.279	94.4
				ML	-0.001	0.219	0.226	94.4
				CML	0.002	0.225	0.233	93.8
D$_1$	BS (Y_1, Z)	opt (Y, Z)	Joint	IPW	0.011	0.228	0.218	96.2
				IPW2	0.011	0.228	0.218	96.2
				ML	0.005	0.204	0.212	93.8
				CML	0.009	0.210	0.216	94.4
D$_2$	-	opt (Y_2, Z)	Marginal	IPW	-	0.221	-	-
				ML	-	0.184	-	-
				CML	-	0.184	-	-
			$E[\phi(X,Z)] = 4$					
A	-	BS (Y_2, Z)	Marginal	IPW	0.037	0.351	0.364	94.2
				ML	0.023	0.299	0.307	94.5
				CML	0.023	0.298	0.307	94.5
B	BS (Y_1, Z)	BS (Y_2, Z)	Marginal	IPW	0.073	0.241	0.241	93.9
				ML	0.092	0.207	0.205	93.4
				CML	0.092	0.207	0.205	93.3
C	BS (Y_1, Z)	BS (Y, Z)	Joint	IPW	0.002	0.270	0.273	94.6
				IPW2	-0.002	0.268	0.275	94.4
				ML	0.001	0.218	0.221	94.2
				CML	0.001	0.223	0.227	94.0
D$_1$	BS (Y_1, Z)	opt (Y, Z)	Joint	IPW	0.007	0.228	0.219	96.0
				IPW2	0.001	0.229	0.225	95.5
				ML	0.001	0.202	0.212	93.2
				CML	0.003	0.208	0.215	94.3
D$_2$	-	opt (Y_2, Z)	Marginal	IPW	-	0.221	-	-
				ML	-	0.184	-	-
				CML	-	0.184	-	-

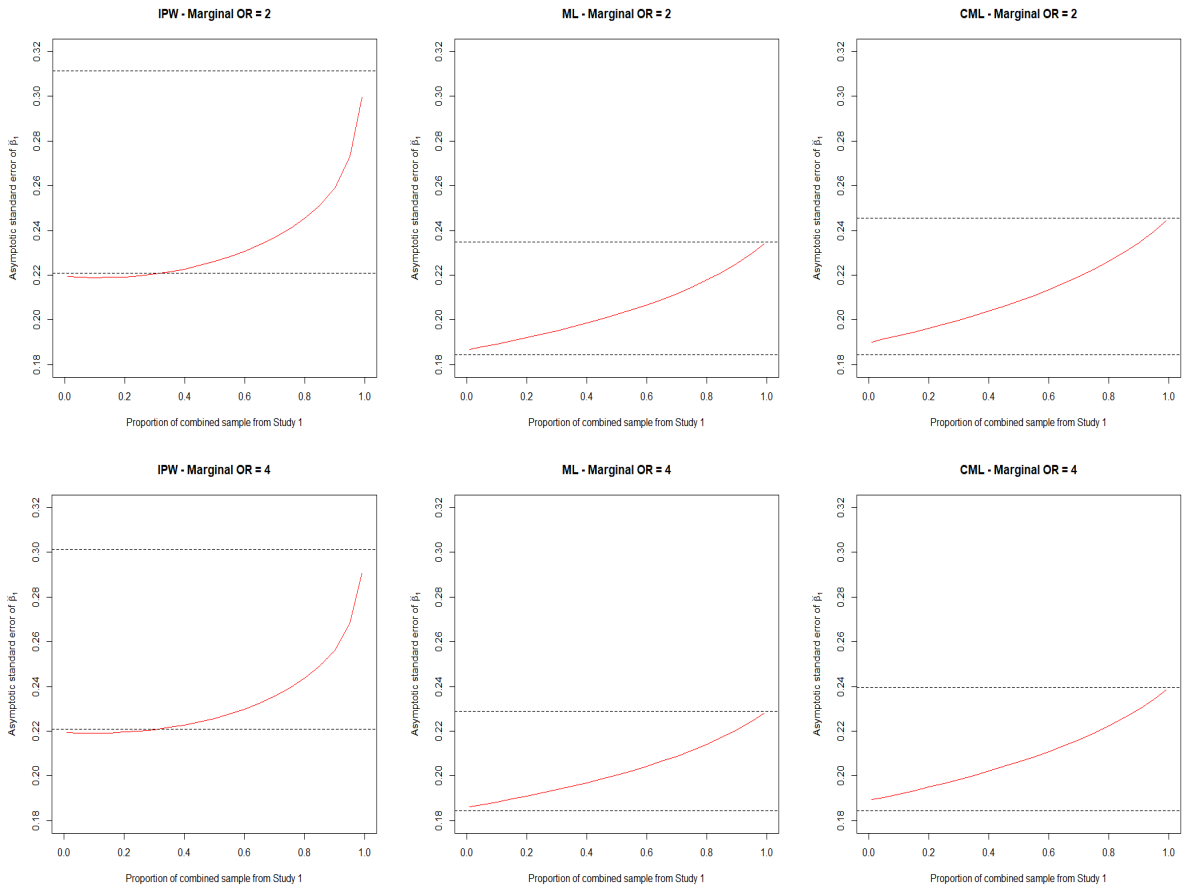

Figure 4.2: Plots of asymptotic standard error of $\hat{\beta}_1$ of designs $\mathbf{D_1}$ as the proportion of the individuals in the combined phase II sub-sample that are selected from Study 1, $E(M_1)/E(M_1+M_2)$, increases. The two rows display graphs of the set-ups with marginal odds ratio (OR) 2 and 4, respectively. Columns from left to right display graphs of frameworks of analysis IPW, ML, and CML, respectively. Lower bounds represent the ideal designs $\mathbf{D_2}$ which select an optimal sub-sample of M_1+M_2 individuals in Study 2. Upper bounds represent using BS to select M_1+M_2 individuals in Study 1. $nsim = 1000$, $n = 5000$, $E(M_1+M_2) = 500$.

Table 4.3: Sampling probabilities of 8 strata defined by (Y_1, Y_2, Z) of our proposed optimal designs $\mathbf{D_1}$ and the ideal optimal designs $\mathbf{D_2}$. The Study 1, Study 2, and Net Study 2 rows refer to π_1, π_2, and $\bar{\pi}_2$ of the proposed designs $\mathbf{D_1}$, respectively. Moderate and strong associations between the responses are reflected by marginal odds ratios 2 and 4, respectively. $nsim = 1000$, $n = 5000$, and $\beta_1 = 0.916$.

Analysis	Designs	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,0)	(1,1,1)
					$E[\phi(X,Z)] = 2$				
Expected Strata Size		2719	574	574	577	131	131	134	161
					Selection Probabilities				
IPW	$\mathbf{D_2}$	0.063	0.063	0.206	0.092	0.092	0.281	0.206	0.281
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.033	0.032	0.170	0.016	0.016	0.089	0.167	0.086
	$\mathbf{D_1}$ Net Study 2	0.051	0.118	0.186	0.103	0.226	0.169	0.240	0.282
ML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.199	0.347	0.347	0.001	0.538
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.270	0.487	0.405	0.088	0.637
CML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.187	0.355	0.354	0.001	0.574
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.259	0.493	0.411	0.088	0.665
					$E[\phi(X,Z)] = 4$				
Expected Strata Size		2777	515	515	601	107	107	193	185
					Selection Probabilities				
IPW	$\mathbf{D_2}$	0.063	0.063	0.206	0.092	0.092	0.281	0.206	0.281
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.034	0.034	0.163	0.017	0.017	0.087	0.160	0.085
	$\mathbf{D_1}$ Net Study 2	0.053	0.119	0.179	0.104	0.227	0.168	0.234	0.281
ML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.198	0.347	0.347	0.001	0.538
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.269	0.487	0.404	0.088	0.637
CML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.187	0.354	0.353	0.001	0.572
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.259	0.492	0.410	0.088	0.664

4.4 Robustness and Model Misspecification

The results so far have shown that adopting a joint response model helps incorporate the information of the sequential two-phase studies rigorously. The advantage comes from specifying the association between the responses, which can be at risk of misspecification. We next investigate the robustness issues of the methods.

4.4.1 Robustness Issues of Secondary Analysis in Two-Phase Designs

Here we investigate the sensitivity and robustness of secondary analyses of two-phase studies to misspecification of the joint response model. Under the simulation and parameter configuration of Section 4.2.3, we now consider the case where the interaction terms of the log odds ratio model are omitted. So with the full model having a dependence structure specified in (4.4) we consdier secondary analyses based on the dependence model

$$\log \phi(X, Z) = \gamma_0 + \gamma_1 X + \gamma_2' Z \ . \tag{4.22}$$

A more serious form of misspecification arises from assuming an exchangeable odds ratio wherein $\phi(X, Z)$ is taken to be a scalar, i.e.,

$$\log \phi(X, Z) = \gamma_0.$$

The simulation results are summarized in Table 4.4.

Recall that we use "Joint" and "Marginal" to denote adopting the joint response model $P(Y|X, Z)$ and the marginal model $P(Y_2|X, Z)$ in the secondary analyses, respectively. The labels "Interaction" and "Exchangeable" refer to types of misspecification of the joint response model by omitting the interaction term and assuming an exchangeable dependence structure, respectively. Since the marginal analysis conducted in IPW does not model the association parameters at all, it is not affected by any type of misspecification. As for the joint analysis, the methods are not affected much if only the interaction term is ignored. This is not surprising as omitting an interaction term in our set-up is equivalent

to estimating $\gamma_3 = 1$ as zero. While we expect the problem to be more serious as γ_3 deviates further away from zero, it is not likely for such coefficients of the interaction term to be arbitrarily large in real-life scenarios. Serious misspecification from assuming an exchangeable dependence structure will affect all inferences based on a joint model, among which IPW2 has relatively lower Ebias and better ECP than CML and ML. Note that although the standard errors increase with misspecification, the joint secondary analysis conducted via IPW is still comparable with the corresponding marginal analysis; see the "Joint Exchangeable IPW" rows versus the "Marginal Exchangeable IPW" rows in Table 4.4 for a comparison.

4.4.2 Robustness Issues of Sequential Two-Phase Designs

Next, we investigate the performances of the proposed sequential two-phase designs when the joint response model is misspecified either by omitting an interaction term in the full model of $\log \phi(X, Z)$ as in (4.22) or assuming an exchangeable dependence structure, in both analysis and the design. The simulation results are summarized in Table 4.5. Similar to Section 4.4.1, omitting an interaction term alone does not affect the performances of designs $\mathbf{C}$ or $\mathbf{D_1}$ much. Other than some slight elevation in the standard errors, the EBias and ECP do not deteriorate much. Hence, it does not preclude the modelling of the association parameters for efficiency while maintaining some level of robustness in such cases, especially when the interaction in the dependence structure is not too strong. On the other hand, the designs give biased results when assuming an exchangeable dependence structure in analysis and design unless they are conducted via IPW. However, we note that we can always fit a full dependence structure when the responses are binary. Finally, we note that our proposed designs $\mathbf{D_1}$ have lower standard error estimates than the designs $\mathbf{C}$ regardless of types of misspecification.

Table 4.4: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the parameter estimates of interest from the IPW and IPW2 (Section 4.2.2), ML (Section 4.2.2), and CML (Section 4.2.2) secondary analyses following a two-phase study based on Y_1 (Study 1) employing balanced sampling with $n = 5000$ and $E(M_1) = 250$. "Interaction" and "Exchangeable" under the "Misspecification" column stand for misspecifying the dependence structure by omitting an interaction term and assuming an exchangeable dependence structure, respectively.

Response Model	$E[\phi(X,Z)]$	Misspecification	Analysis	EBias	ASE	ESE	ECP(%)
Joint	2	Interaction	IPW	0.030	0.443	0.460	92.7
			IPW2	0.039	0.473	0.457	93.2
			ML	0.007	0.341	0.348	94.4
			CML	0.008	0.354	0.359	95.9
		Exchangeable	IPW	0.027	0.442	0.461	93.2
			IPW2	0.164	0.395	0.406	93.6
			ML	0.565	0.315	0.373	57.6
			CML	0.371	0.328	0.318	79.9
	4	Interaction	IPW	0.018	0.433	0.442	93.3
			IPW2	0.023	0.477	0.441	92.8
			ML	-0.004	0.333	0.344	93.7
			CML	-0.003	0.346	0.357	94.4
		Exchangeable	IPW	0.013	0.428	0.453	92.7
			IPW2	0.159	0.378	0.390	93.0
			ML	0.554	0.308	0.411	57.7
			CML	0.345	0.318	0.321	81.3
Marginal	2	Interaction	IPW	0.028	0.442	0.461	93.0
			ML	0.386	0.317	0.318	76.4
		Exchangeable	IPW	0.028	0.442	0.461	93.0
			ML	0.386	0.317	0.318	76.4
	4	Interaction	IPW	0.013	0.428	0.454	92.5
			ML	0.357	0.309	0.320	79.1
		Exchangeable	IPW	0.013	0.428	0.454	92.5
			ML	0.357	0.309	0.320	79.1

Table 4.5: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the estimator of parameter of interest from the combined phase II data using likelihood and inverse weighting methods. "BS" and "opt" stand for balanced sampling and optimal sampling, respectively. "Interaction" and "Exchangeable" under the "Misspecification" column stand for misspecifying the dependence structure by omitting an interaction term and assuming an exchangeable dependence structure, respectively. $nsim = 1000$, $n = 5000$, $E(M_1) = E(M_2) = 0.05n$, and $\beta_1 = 0.916$.

	Stratification				Results			
Design	Study 1	Study 2	Misspecification	Analysis	EBias	ASE	ESE	ECP(%)
			$E[\phi(X,Z)] = 2$					
C	BS (Y_1, Z)	BS (Y, Z)	Interaction	IPW	0.004	0.291	0.281	94.4
				IPW2	0.014	0.271	0.277	94.2
				ML	-0.046	0.218	0.224	94.1
				CML	-0.044	0.223	0.230	94.0
			Exchangeable	IPW	0.001	0.275	0.280	93.7
				IPW2	0.132	0.244	0.246	92.7
				ML	0.358	0.202	0.233	57.0
				CML	0.305	0.208	0.219	67.8
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	Interaction	IPW	0.014	0.234	0.218	96.2
				IPW2	0.022	0.226	0.216	96.2
				ML	-0.017	0.203	0.215	92.9
				CML	-0.011	0.210	0.219	94.5
			Exchangeable	IPW	0.011	0.227	0.219	96.2
				IPW2	0.138	0.215	0.205	92.4
				ML	0.426	0.199	0.232	46.2
				CML	0.450	0.207	0.219	43.4
			$E[\phi(X,Z)] = 4$					
C	BS (Y_1, Z)	BS (Y, Z)	Interaction	IPW	0.007	0.271	0.274	94.7
				IPW2	0.014	0.264	0.269	94.3
				ML	-0.043	0.217	0.217	94.7
				CML	-0.042	0.222	0.223	94.1
			Exchangeable	IPW	0.003	0.270	0.273	94.5
				IPW2	0.142	0.238	0.239	92.7
				ML	0.331	0.203	0.240	62.4
				CML	0.274	0.208	0.221	74.5
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	Interaction	IPW	0.011	0.228	0.219	95.8
				IPW2	0.017	0.225	0.218	96.3
				ML	-0.025	0.201	0.213	92.9
				CML	-0.017	0.207	0.219	93.8
			Exchangeable	IPW	0.007	0.227	0.219	96.6
				IPW2	0.146	0.213	0.206	90.0
				ML	0.248	0.202	0.219	75.6
				CML	0.304	0.204	0.219	67.3

4.5 University of Toronto Psoriatic Arthritis Cohort

For an illustration of our proposed methods, we consider a data set of a subset of the full registry of the University of Toronto Psoriatic Arthritis Cohort (UTPAC) comprised of 706 patients. Upon recruitment to the registry patients are under a detailed clinical examination and provide serum samples for storage in a biobank. Following recruitment, patients undergo annually scheduled clinical examinations to record the extent of joint damage. One particular interest lies in modelling the relation between the Human Leukocyte Antigen (HLA) biomarker *HLA-B27* and progression of joint damage. We note that due to budgetary constraints and the desire to preserve the biospecimens for future use, it is not feasible to assay all stored sera. Auxiliary data on gender and the baseline erythrocyte sedimentation rate (ESR) levels, a traditional inflammatory marker, are available for the entire cohort.

We use this registry to represent a sequence of two-phase studies where the response of Study 1 is defined by the increase in the number of clinically damaged joints over two years. Each of 64 joints is examined at clinical visits and graded for the severity of damage with normal joints graded as 0, deformed joints receiving grade 1, joints featuring ankylosis as grade 2, flailing joints graded 3, and joints receiving surgery graded as 4. We let $Y_1 = 1$ if a patient develops two or more clinically damaged joints of grade 1 or higher in two years from a baseline assessment, and $Y_1 = 0$ otherwise. The binary exposure variable $X = 1$ if the HLA biomarker *B27* is present in the biospecimen, and $X = 0$ otherwise. For an auxiliary covariate, we define $Z = 1$ if the individual's baseline ESR is greater than 20 for females and greater than 13 for males (elevated ESR), and $Z = 0$ otherwise. We suppose that in Study 1, the goal was to investigate the impact of the HLA biomarker *B27* on the progression of clinical joint damage, and that based on this study a sub-sample of 100 patients had their biospecimens assayed to investigate $Y_1|X, Z$. For Study 1, the selection model $\pi_1(Y_1, Z)$ was defined based on BS involving 4 strata defined by (Y_1, Z).

We suppose interest now lies in studying the relationship between the HLA biomarker *B27* and an increase in the number of active joints (swollen joints or joints losing range of motion with pain or tenderness) in this registry. We let $Y_2 = 1$ if a patient develops two or more active joints in two years from the baseline with $Y_2 = 0$ otherwise. We first

perform secondary analyses to fit $Y_2|X, Z$ using Y and Z of the entire dataset, as well as X of the 100 patients in the phase II sub-sample of Study 1. Point estimates and standard error estimates of the parameter of interest β_1 are displayed in the top half of Table 4.6. Assuming an exchangeable odds ratio for the responses leads to distinct point estimates between the likelihood and estimating functions approaches. This suggests that we may have misspecified the joint response model. Closer point estimates are obtained when fitting a full model (4.4) to specify the dependence structure of the responses. See the rows under "Secondary analyses - full dependence" in Table 4.6 for details.

We next consider the case in which our resources permit the assay of 100 biospecimens so that the HLA biomarker *B27* can be determined for more individuals in Study 2. For this aim, Study 1 contributes pilot data to help construct an optimal phase II sub-sample. See the bottom half of Table 4.6 for the results of our proposed optimal design $\mathbf{D_1}$ applied to the PsA study. Assuming an exchangeable odds ratio, estimates based on inverse weighting are more conservative, giving odds ratios of 1.528 (95% CI: 0.626, 3.738) and 1.364 (95% CI: 0.568, 3.278) for IPW and IPW2, respectively, while ML and CML estimators are 2.136 (95% CI: 0.958, 4.762) and 1.667 (95% CI: 0.678, 4.099), respectively. The IPW relies on the selection model alone while the likelihood and IPW2 estimators rely on the joint model; misspecification of the joint model may explain the difference. Fitting a full dependence structure (4.4) gives point estimates that are all generally close, with odds ratios 1.505 (95% CI: 0.669, 3.389) for IPW and IPW2, and 1.589 (95% CI: 0.723, 3.494) and 1.644 (95% CI: 0.715, 3.781) for ML and CML, respectively - the weighted estimators are more robust as the odds ratios are closer than those following the likelihood approaches across various dependence structures. See Figure 4.3 comparing the proportions of individuals selected from the 8 strata defined by (Y_1, Y_2, Z) in Study 2 following the likelihood and inverse weighting methods.

Another round of focused simulation studies framed within the PsA research project is available in Appendix 4B. Secondary use of the biomarker matrix metalloproteinase 3 (MMP-3) is considered to facilitate analyses and sequential two-phase designs to investigate the progression of swollen joints of the patients.

Table 4.6: Point estimates ($\hat{\beta}_1$) and standard error estimates (SE($\hat{\beta}_1$)) of the parameter of interest following the secondary analyses (top half) and the combined phase II data of the sequential two-phase designs (bottom half) using likelihood and inverse weighting methods in the PsA study. For secondary analyses $E(M_1) = 100$, and for sequential two-phase studies $E(M_1) = E(M_2) = 100$. "BS" and "opt" stand for balanced sampling and optimal sampling, respectively.

	Stratification			Results	
Response Model/Design	Study 1	Study 2	Analysis	$\hat{\beta}_1$	SE($\hat{\beta}_1$)
Secondary analyses - exchangeable dependence					
Joint	BS (Y_1, Z)		IPW	0.375	0.843
			IPW2	0.495	0.783
			ML	0.604	0.759
			CML	0.614	0.770
Marginal	BS (Y_1, Z)		IPW	0.372	0.843
Secondary analyses - full dependence					
Joint	BS (Y_1, Z)		IPW	0.369	0.801
			IPW2	0.321	0.850
			ML	0.454	0.809
			CML	0.480	0.806
Marginal	BS (Y_1, Z)		IPW	0.372	0.843
Sequential two-phase studies - exchangeable dependence					
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	IPW	0.424	0.455
			IPW2	0.311	0.447
			ML	0.759	0.409
			CML	0.511	0.459
Sequential two-phase studies - full dependence					
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	IPW	0.409	0.414
			IPW2	0.409	0.414
			ML	0.463	0.402
			CML	0.497	0.425

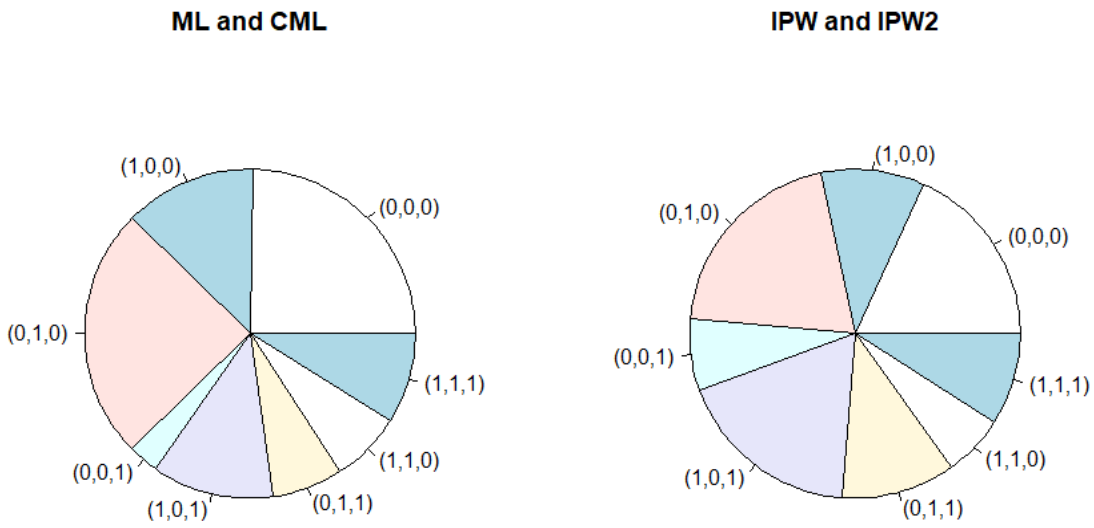

Figure 4.3: Pie charts showing the proportion of individuals selected from 8 strata defined by (Y_1, Y_2, Z) in Study 2 following the optimal designs conducted via likelihood (ML and CML) and inverse weighting (IPW and IPW2) methods. A full model is fitted to specify the dependence of the responses.

4.6 Discussion

We describe and contrast different approaches for the conduct of a secondary analysis of data from two-phase designs. We also consider the design of subsequent two-phase studies when they are based on a common platform. The use of joint response models can enhance the efficiency of estimators based on second-order estimating equations, likelihood and conditional likelihood. Moreover, the estimating functions approach obtains valid estimates under misspecified dependence structures. With the phase II sub-sample size fixed by budgetary constraints, we recommend adopting joint models of the responses with

a full dependence structure and conducting the optimal two-phase designs $\mathbf{D_1}$. By doing so, we can exploit more efficiencies from the likelihood approaches. In particular, the conditional likelihood approach is promising when extending to the settings of continuous exposure variables since it does not model the distribution of the exposures. Should we run out of budget for an upcoming two-phase study, the joint response model further allows us to perform secondary analysis as a last resort.

We present this work in the context of correlated binary responses, but the framework can be adapted to settings where the responses are free to be binary, categorical, continuous, or semicontinuous; with continuous responses, strata may still be defined through the coarsening function by discretization. When secondary responses may be right-censored or interval-censored weighted methods may be more appealing but joint modelling remains a viable approach. Pan et al. (2018b) outlined a secondary analysis of such responses based on IPWEEs in response-dependent sampling designs. Optimal selection models via likelihood approaches are not immediate without making parametric assumptions about the responses. The methods we develop can be extended to deal with the design and analysis of sequential two-phase studies in which the auxiliary covariate is a surrogate of the exposure variables which is appropriate when X is a definitive test result, and Z is an inexpensive, and less accurate test; here $Y \perp Z|X$.

While we consider a common set of auxiliary covariates for the sequence of studies, methods can be generalized to accommodate distinct auxiliary covariates, denoted by Z_1 and Z_2 for Study 1 and 2, respectively. Considering $Z = (Z_1', Z_2')'$ in the joint response model, we can constrain some of the elements of the coefficients in front of Z, i.e., α_2 and β_2, to be 0.

The framework we consider can be naturally extended to deal with an ongoing sequence of several two-phase studies. As shown in Figure 4.1, for a hypothetical Study 3 interested in $E(Y_3|X, Z; \rho)$ following Study 2, one could perform secondary analyses based on $\mathcal{R}_1 \cup \mathcal{R}_2$ using the joint response model $P(Y|X, Z)$ where $Y = (Y_1, Y_2, Y_3)'$. If the aim is to best select M_3 individuals from $\mathcal{R} \setminus (\mathcal{R}_1 \cup \mathcal{R}_2)$ where M_3 reflects its budgetary constraints, then for a given approach to analysis, individuals in $\mathcal{R}_1 \cup \mathcal{R}_2$ serve as pilot data of size $M_1 + M_2$ for preliminary parameter estimates. We approximate the optimal phase II selection model of Study 3 with $\pi_3 = P(R_3 = 1|Y, Z, R_1 = R_2 = 0; \psi_3)$ and the approximately optimal ψ_3

is the one that minimizes

$$\text{asvar}[\sqrt{n}(\hat{\rho}_1-\rho_1);\tilde{\vartheta},\psi_1,\psi_2,\psi_3]+\lambda\left[E(R_3|R_1=R_2=0;\tilde{\vartheta},\psi_1,\psi_2,\psi_3)-\frac{E(M_3)}{n-E(M_1+M_2)}\right],$$

with ϑ augmented to include parameters introduced in Study 3, and ψ_1 and ψ_2 known from Study 1 and Study 2, respectively. If there is concern about specifying the joint response model as the number of studies increases, one could consider modelling the association using conditional pairwise odds ratios to avoid higher-order moments of the responses. Alternatively, one could employ composite likelihood.

Chapter 5

Conclusion and Future Work

The thesis develops methods for the design and analysis of studies involving incomplete data. The key contributions of the earlier chapters are summarized in Section 5.1 and ongoing research topics are considered in Section 5.2. Section 5.3 concludes the thesis by giving a brief outlook of response-dependent sampling with longitudinal data, an important and promising research direction for future research.

5.1 Contributions from Chapters 2 to 4

In Chapter 2, we propose a general framework for implementing the well-known CART algorithm to build survival trees for interval-censored data. We also adapt the framework and evaluate performance for developing survival trees for current status data. Inspired by the imputation technique in Steingrimsson et al. (2019), the framework not only provides a straightforward use of existing software but also avoids relying on common assumptions. We find in simulation studies that our proposed algorithms can make better predictions and better recover the underlying tree structures than *ad hoc* approaches and some alternatives available in the literature. Our proposed method is applied to a medical study of the incidence of axial disease among patients with psoriatic arthritis.

In Chapter 3, we refine the framework of optimal design for two-phase studies proposed

by McIssac and Cook (2015) by considering the estimator following both phase IIA and IIB sub-samples. We extend the work of McIssac and Cook (2015) on inverse probability weighting to examine adaptive two-phase designs conducted via maximum likelihood and conditional likelihood methods; the efficiencies and robustness of the various approaches are then considered. The surrogate value problem is also considered. Such adaptive two-phase designs are well-suited and applied to two-phase biomarker studies in psoriatic arthritis.

In Chapter 4, we focus on large modern cohort studies with biobanks that support biomarker studies. We consider specifically the conduct of a sequence of two-phase studies. We adapt the idea of secondary analysis to the context of two-phase studies and propose the optimal sequential two-phase designs based on information passed on from earlier studies. We propose a joint response model to rigorously incorporate information of the response in a previous study to enhance the efficiency of an upcoming design. Such sequential two-phase studies address the need to use the previously studied biomarkers on a new response while meeting budgetary constraints. The proposed methods are conducted via likelihood approaches and inverse probability weighted estimating equations, followed by an investigation of efficiency and robustness.

We here comment that the theme in Chapter 2 differs from that in Chapters 3 and 4, as they address different motivating concerns involving incomplete data. While missing responses arise from periodic assessments in medical studies, missing covariates by design are mainly attributed to budgetary constraints in observational studies. Despite this, there are potential methodological connections between the themes as we approach complicated problems, such as two-phase studies with interval-censored responses. Indeed, one could stratify the phase I sample based on the assessments or the interval length in such two-phase studies. Still, the imputation techniques discussed in Chapter 2 may help bridge the gap between interval censoring and the optimal designs proposed in Chapters 3 and 4. For example, in the context of an adaptive two-phase design, the phase IIA sub-sample may help construct estimators of the conditional survivor function for response imputations. Stratification and the formation of a class of selection models naturally follow, within which the optimal phase IIB sub-sampling scheme of our interest can be considered.

5.2 Future Research

5.2.1 Ensemble Prediction Methods for Interval-Censored Data

There has been much interest in the development of ensemble prediction methods in the past few decades. Such methods select a collection (ensemble) of prediction models and combine their predictions to improve predictive performance. The prediction is usually based on the majority vote from all the predictions of the individual predictors in the ensemble. Bootstrap Aggregation, often known as bagging (Breiman, 1996), fits multiple CART trees on bootstrapped samples which are obtained by sampling with replacement multiple times from the training data sets. It then averages all the unpruned CART trees fitted on the bootstrap samples. Using bootstrap to assess the variance of the estimators, bagging achieves variance reduction at the expense of tiny increases in bias and hence, becomes particularly useful for estimators with high variances. Another ensemble method is random forests (Breiman, 2001), which further weakens the dependence among the single prediction models in the ensemble in the sense that they do not necessarily consider the same set of splitting variables. Unlike bagging, random forests conduct searches only over a random subset of the variables at each splitting point. As one of the top-performing prediction methods, random forests can currently be implemented in the `R` package `randomForest`.

Recent developments of ensemble prediction methods have been able to accommodate censored responses. Ishwaran et al. (2008) developed the random survival forest for right-censored data using the log-rank test statistics as the splitting criterion. The method was refined by Zhu and Kosorok (2012) as the recursively imputed survival trees. Recent work on the ensemble prediction methods for interval-censored responses includes Cho et al. (2019) which proposed an iterative tree-based ensemble method. The nonparametric regression estimator is obtained by iteratively updating the estimates of the survivor functions, and can be viewed as a self-consistent estimator with convergence monitored using out-of-bag samples. The work is implemented via the `R` package `icrf`. Another advance is in Yao et al. (2019), who extended the work of Fu and Simonoff (2017) to explore the use of ensemble methods based on the conditional inference survival forest.

It is of natural interest to develop the ensemble prediction methods corresponding to our proposed regression trees based on censoring unbiased transformations and pseudo-observations in Chapter 2. The ensemble predictor for interval-censored responses may differ from that for current status data, as the NPMLE of the survivor function has a closed-form expression only for current status data. We expect that such ensemble learners would improve the prediction accuracy of the performance compared to the single tree prediction models.

5.2.2 Current Status Composite Likelihood

We here propose a current status composite likelihood as an alternative to build survival trees for interval-censored data. Suppose that there is a type K interval censoring observation process, and the visit process is completely independent of the failure times. Let $u_{i0}, \ldots, u_{iR_i}$ denote the assessments for individual $i = 1, \ldots, n$. Let $Y_i(\cdot)$ denote the binary response of the observations so that $Y_i(u_{ij}) = 1$ if individual i is observed to have failed at assessment u_{ij}, $j = 1, \ldots, R_i$, and $Y_i(u_{ij}) = 0$ otherwise. Adopting the likelihood for current status data, we have

$$L = \prod_{i=1}^{n} \prod_{j=0}^{R_i} S_i(u_{ij}|X_i)^{1-Y_i(u_{ij})}[1 - S_i(u_{ij}|X_i)]^{Y_i(u_{ij})},$$

where S denotes the conditional survivor function, and X the covariates that are assumed to be not time-dependent. Such likelihood employs the idea of pseudo individuals in the sense that the status at visit $j = 1, \ldots, R_i$ can be considered as contributions from distinct individuals. As a result, R_i pseudo individuals arise, and a composite likelihood in which information is repeatedly used follows. As the survivor function for current status data is easier to estimate, the resulting composite likelihood may lead to faster implementation compared to the ordinary likelihood for interval-censored data. While the prediction performance of such models warrants checking, it is of our interest to inspect the loss of efficiency of such composite likelihood compared to maximum likelihood. The problem is more complicated if there is a need to model the visit process; Jiang and Cook (2020) used such composite likelihood in the context of clustered multistate processes under intermittent observation with aggregation.

5.2.3 Two-Phase Designs via Calibration

Recent work on two-phase designs includes the calibrated inverse probability weighted (IPW) estimator that enhances statistical efficiency (Rivera-Rodriguez et al., 2019). Adopting the IPW framework of analysis, the aim is to use calibration as a means to improve the efficiency of the estimator. This is achieved by finding the so-called "calibrated weights", $\tilde{w}$, that minimizes certain distances to the original weights w used in the inverse weighting analysis subject to the calibration constraints. Rivera-Rodriguez et al. (2019) suggested the distance function as the χ^2 distance, $d(\tilde{w}, w) = (\tilde{w} - w)^2 / 2w$, or the deviance distance, $d(\tilde{w}, w) = \tilde{w} \log(\tilde{w}/w) - \tilde{w} + w$. As for the calibration constraints, Rivera-Rodriguez et al. (2019) suggested the influence functions of the mean model, which are the columns of the matrix

$$X_{IF} = -U(\beta) E \left[\frac{\partial}{\partial \beta} U(\beta) \right]^{-1},$$

where $U(\beta)$ is the estimating equations to solve for the parameter of interest β. The more efficient calibrated IPW estimator is obtained by replacing w with $\tilde{w}$ in the analysis. The procedure makes use of information available in phase I that is not involved in the design and hence, enhances statistical efficiencies of the two-phase designs. In survey sampling, such calibration helps exploit the auxiliary information to improve the Horvitz-Thompson estimator (Lumley et al., 2011; Deville and Sardnal, 1992). Chen and Lumley (2020) used the term "generalized ranking" to describe the more efficient class of design-based estimators. As a result, one of the extensions of our work on two-phase designs is to compare the calibrated IPW estimator to those following from maximum likelihood and conditional maximum likelihood approaches. The potential link between the model-based estimation methods discussed in Chapter 3 and the design-based estimation methods in Chen and Lumley (2020) warrant attention.

5.2.4 Two-Phase Designs in Longitudinal Settings

Modern epidemiological cohort studies involve the longitudinal collection of responses from individuals over time. The Canadian Longitudinal Study of Aging, for example, is a national research initiative in which approximately 30,000 recruited individuals will be fol-

lowed for up to 21 years to record data on health outcomes (Raina et al., 2009). Serum samples are obtained upon recruitment, and detailed clinical examinations are scheduled to take place every three years. Many health research projects based on this unique infrastructure aim to investigate the association between biomarkers measurable in the stored serum samples and longitudinal responses, but measurement of biomarkers for all participants is cost-prohibitive. The need to preserve biospecimens and control costs has naturally led to the development of response-dependent two-phase sampling designs. Much of this work on two-phase designs has been directed at cross-sectional or retrospective data. McIssac and Cook (2013) investigated two-phase sampling designs in the context of studies with repeated measurements and a progressive process. Recent work in the area includes likelihood-based analysis for fitting generalized linear mixed models to longitudinal data from response-related sampling designs (Neuhaus et al., 2014). Conditioning on the sampling protocol and summary statistics of the longitudinal responses, Neuhaus et al. (2014) modified the standard conditional likelihood approach to incorporate random intercept models with a canonical link. Alternative approaches include two-phase stratified sampling designs in which distinct strata are created from subject-specific response summaries (Schildcrout et al., 2013, 2018), or by low-dimensional numerical summaries of longitudinal responses and confounders (Schildcrout et al., 2019), both followed by identifying a highly informative sub-cohort. Such sampling designs appear to be highly efficient relative to random sampling. Haneuse and Rivera (2018) discussed a general approach to analyze cluster-correlated case-control data based on inverse probability weighted estimating equations (IPWEE). Amorim (2019) further investigated semiparametric estimators for secondary analyses with correlated responses and proposed an estimated conditional maximum likelihood method.

Two-phase designs with longitudinal data represent an exciting research area with numerous areas worthy of development. In the retrospective setting in which individuals are selected for biomarker measurements upon the completion of the followup, budgetary constraints limit the number of biospecimens that can be assayed. If the biomarker of interest is not time-dependent, then adaptive two-phase designs in the spirit of Chapter 3 will be applicable to maximize statistical efficiency. Tao et al. (2021) also proposed two-wave two-phase response-dependent sampling designs addressing this topic. Adapting the ascer-

tainment corrected maximum likelihood approach (Schildcrout and Heagerty, 2011), Tao et al. (2021) estimated the conditional distribution of the unobserved exposure variables followed by applying a multiple imputation procedure (Rubin, 1987) for analysis. The optimal design was achieved via a thorough search in the design space. In Section 5.3.1, we layout two-phase designs with retrospective longitudinal data based on IPWEEs.

The problem becomes more complicated with time-dependent biomarkers since it may be desirable to sample an individual at more than one time point. For longitudinal binary response data, Schildcrout and Heagerty (2008) and Schildcrout et al. (2018) recommended the phase II sub-sampling to exclusively select individuals who experience the event at some but not all time points. We provide a brief sketch of the two-phase designs involving a joint model of the selection indicators at the end of Section 5.3.1.

If sub-sampling is to be carried out in a serial prospective fashion, individuals could be selected for biomarker measurements at time point k based on the responses and auxiliary covariates up to time point k, as well as the observed biomarkers measured up to time point $k-1$. The work on sequential two-phase designs in Chapter 4 builds a foundation for such longitudinal studies with a time-fixed exposure and serial budgetary constraints. As discussed in Section 4.6, it is natural to extend Study 1 and Study 2 to additional sequential two-phase studies, each targeting those who remain eligible after the earlier studies. Upon completion of the entire followup, all phase II sub-samples will be combined for analysis. If the biomarker of interest is not time-dependent, then no individuals shall be assessed more than once throughout the longitudinal study. Therefore, the assessments can be regarded as a sequence of two-phase studies conducted on the same platform, where the longitudinal responses can be treated as the sequential responses discussed in Chapter 4. Moreover, the sampling constraints in Chapter 4 reflect the budgetary limitations of the repeated examinations, and the selected models of the sequential two-phase studies can be adapted to the prospective longitudinal setting. While we anticipate analyses to be more challenging as the number of studies increases, the parameter of interest is usually identical in a longitudinal setting. As for time-varying exposures in serial prospective studies, we need to address not only the serial budgetary constraints but also the correlation among the selection indicators. See Section 5.3.2 for a brief layout of such two-phase designs.

5.3 Two-Phase Studies with Longitudinal Data

5.3.1 Retrospective Two-Phase Sampling with Longitudinal Data

Here we propose an adaptive two-phase design with retrospective longitudinal data in which subjects are selected for biomarker measurements such that the resultant estimators have maximum efficiency subject to budgetary constraints. The optimal design addresses the serial dependence of the longitudinal responses over time.

We consider a study involving a sample of n independent individuals assessed at each of K time points where Y_{ik} is the response at assessment k for individual i, $k = 1, \ldots, K$. The complete response vector is $Y_i = (Y_{i1}, \ldots, Y_{iK})'$, $i = 1, \ldots, n$. Let X_{i1} denote an incompletely observed covariate reflecting a biomarker of interest, $X_{i2} = (X_{i21}, \ldots, X_{i2p})'$ a $p \times 1$ auxiliary covariate vector which is always observed, and $X_i = (X_{i1}, X_{i2}')'$, $i = 1, \ldots, n$. Assuming binary responses and time-fixed binary covariates, suppose interest lies in the marginal mean $\mu_{ik} = E(Y_{ik}|X_i)$ such that

$$\text{logit}\mu_{ik} = \beta_0 + \beta_1 X_{i1} + \beta_2' X_{i2}, \tag{5.1}$$

where β_1 is the parameter of primary interest characterizing the association between the expensive biomarker and the response given the auxiliary covariates, and $E(Y_i|X_i)$ is the $K \times 1$ vector $\mu_i = (\mu_{i1}, \ldots, \mu_{iK})'$. The vector $\beta = (\beta_0, \beta_1, \beta_2')'$ has dimension $(1+1+p) \times 1$. The conditional association of the binary responses over assessments is characterized by pairwise log odds ratios

$$\alpha_{ijk} = \log \text{OR}(Y_{ij}, Y_{ik}|X_i) = \log \frac{P(Y_{ij} = 1, Y_{ik} = 1|X_i)/P(Y_{ij} = 0, Y_{ik} = 1|X_i)}{P(Y_{ij} = 1, Y_{ik} = 0|X_i)/P(Y_{ij} = 0, Y_{ik} = 0|X_i)}, \tag{5.2}$$

for $j < k, j, k = 1, \ldots, K$. Under an exchangeable dependence structure, $\alpha = \alpha_{ijk}$ is a scalar, but more generally we consider α as a vector of length $L = K(K-1)/2$. We then let $\theta = (\alpha', \beta')'$ of dimension $L + 1 + 1 + p$ denote the parameters of the response model. We also let $g(X_i; \gamma) = g_1(X_{i1}|X_{i2}; \gamma_1) g_2(X_{i2}; \gamma_2)$, $\vartheta_1 = (\theta', \gamma_1')'$, and $\vartheta = (\theta', \gamma')'$.

The phase I sample is comprised of n individuals giving data $\{Y_i, X_{i2} : i = 1, \ldots, n\}$ with X_{i1} unknown. A phase II sub-sample of $M < n$ biospecimens is chosen and assayed

with X_{i1} recorded for the selected individuals. In an ordinary two-phase design, a selection indicator $R_i = 1$ if individual i is selected for the phase II sub-sample and $R_i = 0$ otherwise. See Table 5.1 for an illustration of such data comprised of n individuals. Adopting notation in Chapter 3, we consider an adaptive two-phase design in which the phase II sub-sample is comprised of a phase IIA sub-sample selected with a convenient sampling strategy for a preliminary estimate of ϑ and a phase IIB sub-sample constructed to maximize the precision of estimator of β_1 from the entire phase II sub-sample. Recall that we let A denote the selection indicator of phase IIA which is realized based on selection model $\pi_A(\psi_A)$. Likewise, we let B denote the selection indicator of phase IIB which is realized based on selection model $\pi_B(\psi_B)$. While phase IIA selects M_A out of n individuals, phase IIB selects M_B out of $n - M_A$ individuals that remain eligible.

Table 5.1: An illustration of two-phase data of n individuals.

Response	Y_1	Y_2	..	Y_i	..	Y_n
Biomarkers	X_{11}	?	..	?	..	X_{n1}
Auxiliary Covariates	X_{12}	X_{22}	..	X_{i2}	..	X_{n2}
Selection Indicators	$R_1 = 1$	$R_2 = 0$	..	$R_i = 0$	..	$R_n = 1$

Analysis Methods and Associated Design

Such adaptive two-phase designs can be based on second-order IPWEEs that give a marginal formulation of covariate effects modelling both the mean and the association parameters (Zhao and Prentice, 1990). Following (5.2), the odds ratio can be expressed as

$$\mathrm{OR}(Y_{ij}, Y_{ik}|X_i) = \exp(\alpha_{ijk}) = \frac{\eta_{ijk}(1 - \mu_{ij} - \mu_{ik} + \eta_{ijk})}{(\mu_{ij} - \eta_{ijk})(\mu_{ik} - \eta_{ijk})},$$

where $\eta_{ijk} = E(Y_{ij}Y_{ik}|X_i)$, $j < k, j, k = 1, \ldots, K$. Here η_{ijk} is a function of means μ_{ij}, μ_{ik}, and the pairwise log odds ratio α_{ijk} using the quadratic formula. Moreover, the $K \times K$ covariance matrix $\mathrm{Cov}(Y_i|X_i)$ with kth diagonal element $\mathrm{Var}(Y_{ik}|X_i) = \mu_{ik}(1-\mu_{ik})$ for $k = 1, \ldots, K$ and (j, k) off-diagonal element $\mathrm{Cov}(Y_{ij}, Y_{ik}|X_i) = \eta_{ijk} - \mu_{ij}\mu_{ik}$ for $j < k, j, k = 1, \ldots, K$ is indexed by θ.

Inverse Probability Weighted Estimating Equations

Let $W_i = (W_{i1}, \ldots, W_{iL})' = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i(K-1)}Y_{iK})'$ denote the pairwise products that can be formed among the longitudinal responses. In particular, for individual i, $W_{ijk} = Y_{ij}Y_{ik}$ for $j < k$, $j, k = 1, \ldots, K$. Let $Z_{i1} = (Y_i', W_i')'$ denote the $(K + L) \times 1$ augmented response vector with $E(Z_{i1}|X_i) = \xi_{i1}$. The phase IIA parameter estimates can be obtained by solving the second-order IPWEE

$$\sum_{i=1}^{n} U_{i1A}(Y_i|X_i; \theta, \psi_A) = \sum_{i=1}^{n} D_{i1}' \Sigma_{i1}^{-1} \frac{A_i}{\pi_{iA}} (Z_{i1} - \xi_{i1}) = 0, \tag{5.3}$$

where $D_{i1} = \partial \xi_{i1}/\partial \theta$ is a $(K + L) \times (L + 1 + 1 + p)$ matrix and $\Sigma_{i1} = \text{Cov}(Z_{i1}|X_i)$ is a $(K + L) \times (K + L)$ working matrix (Robins et al., 1995; Prentice and Zhao, 1991). While the top left $K \times K$ sub-matrix of Σ_{i1} has been fully specified with diagonal elements $\mu_{ik}(1 - \mu_{ik})$ and off-diagonal elements $\eta_{ijk} - \mu_{ij}\mu_{ik}$ for $j < k, j, k = 1, \ldots, K$, the remaining elements involve third and fourth moments of the longitudinal responses. We adopt the block diagonal working matrix which specifies the bottom right $L \times L$ sub-matrix to be diagonal with elements

$$\text{Var}(W_{ijk}|X_i) = E(Y_{ij}^2 Y_{ik}^2|X_i) - E(Y_{ij}Y_{ik}|X_i)^2 = \eta_{ijk} - \eta_{ijk}^2$$

for $j < k, j, k = 1, \ldots, K$ and zero otherwise to avoid computing higher-order moments of Y_i. The nuisance parameter γ is separately estimated as in Section 3.2. Upon completion of the phase IIB sub-sampling, the IPWEEs that marginalize over the selection process are

$$\sum_{i=1}^{n} U_{i1B}(Y_i|X_i; \theta, \bar{\psi}^*) = \sum_{i=1}^{n} \frac{\bar{R}_i^*}{\bar{\pi}_i^*} D_{i1}' \Sigma_{i1}^{-1} (Z_{i1} - \xi_{i1}) = 0 \tag{5.4}$$

and

$$\sum_{i=1}^{n} U_{i2B}(A_i, B_i; \bar{\psi}^*) = \sum_{i=1}^{n} \frac{\partial \bar{\pi}_i^*}{\partial \bar{\psi}^*} \frac{1}{\bar{\pi}_i^*(1 - \bar{\pi}_i^*)} (\bar{R}_i^* - \bar{\pi}_i^*) = 0, \tag{5.5}$$

where $\bar{R}_i^* = A_i + (1 - A_i)B_i$ and $\bar{\pi}_i^* = \pi_{iA} + (1 - \pi_{iA})\pi_{iB}$ indexed by $\bar{\psi}^*$. Under regularity conditions, the resulting estimator $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \Gamma_B^{-1}(I_B - H_B \Omega_B H_B')(\Gamma_B^{-1})'), \text{ as } n \to \infty, \tag{5.6}$$

where $\Gamma_B = E[-\partial U_{i1B}(Y_i|X_i;\theta,\bar{\psi}^*)/\partial\theta]$, $I_B = E[U_{i1B}(Y_i|X_i;\theta,\bar{\psi}^*)U'_{i1B}(Y_i|X_i;\theta,\bar{\psi}^*)]$, $\Omega_B = E[U_{i2B}(A_i,B_i;\bar{\psi}^*)U'_{i2B}(A_i,B_i;\bar{\psi}^*)]$, and $H_B = E[-\partial U_{i1B}(Y_i|X_i;\theta,\bar{\psi}^*)/\partial\bar{\psi}^*]$ (Robins et al., 1995). Note that the asymptotic covariance matrix in (5.6) has size $(L+1+1+p)\times(L+1+1+p)$. Focusing on the (β_1,β_1) component, the approximately optimal ψ_B is the one that minimizes the Lagrangian (3.2).

Likelihood and Conditional Likelihood

Alternatively, such adaptive two-phase designs with longitudinal responses can be conducted via likelihood analyses. With longitudinal data and the desire for robust inference, this is perhaps less appealing than the settings of Chapters 3 and 4 but we consider it here for completeness. If ϑ and ψ are functionally independent, then the partial likelihood

$$L(\vartheta) \propto \prod_{i=1}^{n}[P(Y_i|X_i)g_1(X_{i1}|X_{i2})]^{A_i}\left[\int P(Y_i|x_1,X_{i2})g_1(x_1|X_{i2})dx_1\right]^{1-A_i} g_2(X_{i2}),$$

helps obtain the phase IIA parameter estimates. Similarly, considering the entire phase II sub-sample, the score equation for estimating ϑ_1 is

$$S_B(Y,X_1|X_2;\vartheta_1) = \sum_{i=1}^{n} S_{iB}(Y_i,X_{i1}|X_{i2};\vartheta_1) = \sum_{i=1}^{n}\begin{pmatrix} S_{i1B}(\vartheta_1) \\ S_{i2B}(\vartheta_1)\end{pmatrix} = 0$$

of the form (3.3). Under regularity conditions, asymptotic normality follows, and the optimal ψ_B is the one that minimizes (3.2).

The score equation of the conditional likelihood approach similarly follows from (3.8), while the optimal ψ_B minimizes (3.9). The only caveat is that we have suppressed the time index in the score equations. Contrary to a standard adaptive two-phase design presented in Chapter 3, the response model for longitudinal responses is $P(Y_i|X_i) = P(Y_{i1},\ldots,Y_{iK}|X_i;\theta)$ for $i = 1,\ldots,n$. For computational purposes, we can adopt reasonable assumptions for simplification, for example, assuming Markov property so that $Y_k \perp Y_j|Y_{k-1}$ for $j < k-1$, $k = 1,\ldots,K$. We also point out the possibility of using composite likelihood, especially when K is large. In such cases, the asymptotic covariance matrix has the sandwich form.

Empirical Studies

Here follows a preliminary round of simulations of adaptive two-phase designs with retrospective longitudinal data and time-fixed covariates. We consider a longitudinal study with three time points $K = 3$, sample size $n = 5000$, phase II sub-sample of expected size $E(M) = 1500$ chosen via Bernoulli sampling, and $nsim = 1000$. We assumed that the longitudinal responses, biomarkers, auxiliary covariates, and the selection indicators are scalar binaries, i.e., $p = 1$. For each individual, the longitudinal responses were assumed to have an exchangeable dependence structure, i.e., a scalar pairwise odds ratio. The binary covariates satisfied (3.16) and (3.17). The parameter configuration $(\beta_0, \beta_1, \beta_2, \alpha, \gamma_2, \gamma_{10}, \gamma_{11}) = (-1.71, \log 2.5, \log 1.5, \log 2, -1.39, -1.75, \log 4)$ ensures that the marginal probabilities of both the responses and covariates are 0.2, while a strong association presents among the responses, biomarker, and the auxiliary covariate. The selection model in phase IIA and phase IIB are characterized as

$$\text{logit}\pi_A(Y, X_2; \psi_A) = \psi_{A0} + \psi_{A1} I(Y_\cdot > 0) + \psi_{A2} X_2 + \psi_{A3} I(Y_\cdot > 0) X_2;$$

$$\text{logit}\pi_B(Y, X_2; \psi_B) = \psi_{B0} + \psi_{B1} I(Y_\cdot > 0) + \psi_{B2} X_2 + \psi_{B3} I(Y_\cdot > 0) X_2,$$

indexed by $\psi_A = (\psi_{A0}, \psi_{A1}, \psi_{A2}, \psi_{A3})'$ and $\psi_B = (\psi_{B0}, \psi_{B1}, \psi_{B2}, \psi_{B3})'$, respectively. In other words, the phase I sample is divided into four strata according to the binaries $I(Y_\cdot > 0)$ and X_2, where $Y_\cdot = \sum_{i=1}^{3} Y_i$.

The longitudinal responses were generated from the conditional linear family (Qaqish, 2003), which is defined as a sub-family of the multivariate binary distributions with a given mean and covariance. For $i = 2, \ldots, n$, Qaqish (2003) showed that the conditional distribution of the correlated responses satisfy

$$E(Y_i | Y_1, \ldots, Y_{i-1}) = \mu_i + \sum_{j=1}^{i-1} \sum_{a \in A_{ij}} K_{ij}(a) \prod_{t=1}^{j} (Y_{a_t} - \mu_{a_t}),$$

where A_{ij} is the set of integer j vectors $\{a : 1 \leq a_1 < \ldots < a_j \leq i-1\}$, and K_{ij} is a column vector of length $i-1$ chooses j. The conditional linear family is obtained by setting $K_{ij} = 0$ for $i = 2, \ldots, n$ and $j = 2, \ldots, i-1$. In particular, when the responses have equal

means and an exchangeable correlation, explicit formula for the conditional expectation is available. In our set-up, conditioning on X,

$$E(Y_i|Y_1,\ldots,Y_{i-1},X_i) = \frac{(1-\rho)\mu_i + \rho\sum_{j=1}^{i-1} Y_j}{1+\rho(i-2)}$$

for $i = 2, 3$, where ρ refers to the correlation of the responses (Preisser et al., 2002). Alternatives of generating correlated binary random variables include the multivariate probit method (Emrich and Piedmonte, 1991). See Preisser and Qaqish (2014) for a detailed review and comparison between the methods.

To demonstrate the efficiency gains, the adaptive two-phase designs were compared to the standard non-adaptive designs. As shown in Table 5.2, the adaptive two-phase designs based on the IPWEE framework of analysis for longitudinal responses are more efficient than the standard SRS and BS designs. The average standard errors (ASE) match the empirical standard errors (ESE), and the empirical coverage probabilities (ECP) are compatible with the nominal 95% levels for all phase IIA sub-sample sizes.

Table 5.2: Average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of estimators from second-order IPWEE (IPW) adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.2E(M)$ or $0.50E(M)$. Non-adaptive SRS and BS designs are included as the "100% IIA" columns. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 0.3n$, and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

					Proposed Adaptive Designs					
		100% IIA			50% IIA			20% IIA		
Analysis	IIA	ASE	ESE	ECP%	ASE	ESE	ECP%	ASE	ESE	ECP%
IPW	SRS	0.107	0.106	95.0	0.098	0.099	95.6	0.098	0.094	95.0
	BS	0.109	0.109	94.9	0.099	0.096	94.6	0.098	0.098	94.5

Extensions to Accomodate Time-Varying Exposures

Should we have a time-dependent biomarker, let X_{i1k} denote the expensive biomarker at assessment k. In such cases, we assume that the marginal mean $\mu_{ik} = E(Y_{ik}|X_{i1k}, X_{i2})$ and the pairwise log odds ratio $\alpha_{ijk} = \log \text{OR}(Y_{ij}, Y_{ik}|X_{i1j}, X_{i1k}, X_{i2})$ still has the form of (5.1) and (5.2), respectively, except with the time-dependent covariates updated accordingly. As a result, the availablity of the biomarker for individual i will be indicated by a $K \times 1$ vector, $R_i = (R_{i1}, \ldots, R_{iK})'$. Let the marginal selection model be $\pi_{ik} = P(R_{ik} = 1|Y_i, X_{i2}; \psi_k)$. The longitudinal association of the selection indicators can be characterized by pairwise log odds ratios

$$\omega_{ijk} = \log \frac{P(R_{ij} = 1, R_{ik} = 1|Y_i, X_{i2})/P(R_{ij} = 0, R_{ik} = 1|Y_i, X_{i2})}{P(R_{ij} = 1, R_{ik} = 0|Y_i, X_{i2})/P(R_{ij} = 0, R_{ik} = 0|Y_i, X_{i2})}, \tag{5.7}$$

where ω is a vector of length L. Let

$$Q_i = (Q_{i1}, \ldots, Q_{iL})' = (R_{i1}R_{i2}, \ldots, R_{i(K-1)}R_{iK})'$$

denote the pairwise products formed among the selection indicators with

$$\zeta_{ijk} = E(R_{ij}R_{ik}|Y_i, X_{i2}) = P(R_{ij} = R_{ik} = 1|Y_i, X_{i2}),$$

and $Z_{i2} = (R_i', Q_i')'$ the augmented selection vector with $E(Z_{i2}|Y_i, X_{i2}) = \xi_{i2}$ indexed by both ω and $\psi = (\psi_1', \ldots, \psi_K')'$. The second-order IPWEE for the responses, therefore, involves a weighting matrix instead of a single weight. Specifically, one could consider

$$\sum_{i=1}^{n} U_{i1}(Y_i|X_i; \theta, \psi, \omega) = \sum_{i=1}^{n} \Delta_i D_{i1}' \Sigma_{i1}^{-1}(Z_{i1} - \xi_{i1}) = 0, \tag{5.8}$$

where

$$\Delta_i = \begin{pmatrix} \text{diag}\left(\frac{R_{ik}}{\pi_{ik}(Y_i, X_{i2})}\right) & 0 \\ 0 & \text{diag}\left(\frac{R_{ij}R_{ik}}{\zeta_{ijk}(Y_i, X_{i2})}\right), \end{pmatrix} \tag{5.9}$$

$j < k$, $j, k = 1, \ldots, K$, is a block diagonal matrix with blocks of size $K \times K$ and $L \times L$. As for the selection model, one could consider

$$\sum_{i=1}^{n} U_{i2}(R_i; \psi, \omega) = \sum_{i=1}^{n} D_{i2}' \Sigma_{i2}^{-1}(Z_{i2} - \xi_{i2}) = 0, \tag{5.10}$$

where $D_{i2} = \partial \xi_{i2} / \partial (\psi', \omega')$ and $\Sigma_{i2} = \text{Cov}(Z_{i2}|Y_i, X_{i2})$ is a block diagonal working matrix. Such IPWEEs allow us to perform analyses of an adaptive two-phase design; assigning the vector of selection indicators R_i as A_i and $\bar{R}_i^* = A_i + (1 - A_i)B_i$, respectively, followed by updating the corresponding selection models π and ζ in (5.8) and (5.10) will do. The approximately optimal phase IIB sub-sampling follows from minimizing the asymptotic variance of the estimator of β_1 obtained by the IPWEEs subject to budgetary constraints.

Assumptions on the sampling scheme can simplify the analysis presented above. For example, if we decide to either fully observe the covariates of individual i or not to observe X_{i1} at any point, then the components of R_i are identical. Another possible simplification comes from independent selections of assessments for individuals (i.e., we could set $\zeta_{ijk} = \pi_{ij}\pi_{ik}$). Both assumptions greatly simplify the weighting matrix Δ and the covariance matrix Σ_2. However, the IPWEEs (5.8) and (5.10) address the general situation where the selections of individuals are allowed to correlate across assessments.

Multi-Dimensional Biomarkers of Interest

Finally, we comment on the possibility and associated challenges of dealing with a multi-dimensional biomarker of interest, i.e., $X_{i1} = (X_{i11}, \ldots, X_{i1q})'$, $i = 1, \ldots, n$. For individuals selected in the phase II sub-sampling, their biospecimens are assayed, and the full vector of X_{i1} is recorded. If the biomarker of interest is time-dependent, then X_{i1k} is a $q \times 1$ vector at assessment $k = 1, \ldots, K$. We expect the analyses and design to be more challenging when the biomarker of interest is multi-dimensional ($q > 1$) compared to our discussion above ($q = 1$) for the following reasons. The biomarker distribution involves the specification of a multivariate model that is harder to estimate and work with when q is large. As a result, having multi-dimensional biomarkers of interest affects both the maximum likelihood analysis and the designs for all approaches since they all require modelling $X_1|X_2$. Furthermore, the objective function in Lagragian (3.2) becomes unclear when β_1 is a vector. One could consider minimizing the trace of the (β_1, β_1) submatrix of the asymptotic covariance matrix.

5.3.2 Prospective Longitudinal Two-Phase Sampling

Time-Fixed Exposures

Another longer-term scientific research topic is two-phase studies with serial prospective longitudinal data subject to multiple constraints over time. Given that the covariates are time-fixed, such prospective studies with K assessments can be considered as K sequential two-phase studies in the context of Chapter 4 with identical parameter of interest. Adopting notation in Section 5.3.1, we use logistic models and conditional pairwise odds ratios to characterize the joint response model of the longitudinal responses $Y = (Y_1, \ldots, Y_K)'$. For convenience, we clarify the following notation for the rest of the section. We let a capital letter by itself denotes a vector quantity throughout K assessments, and a subscript k denotes the quantity at assessment k, $k = 1, \ldots, K$. We use an additional macron to denote the history of the quantity up to assessment k. For example, $R = (R_1, \ldots, R_K)'$ where R_k is the selection indicator at assessment k, $1, \ldots, K$. The history of the selection indicator up to assessment k is then $\bar{R}_k = (R_1, \ldots, R_k)'$. Moreover, we use a superscript star to denote a combination of the history of the quantity up to assessment k whenever required. For example, $\bar{R}_2^* = R_1 + (1 - R_1)R_2$ denotes the selection indicator of the net selection model up to the second assessment. Suppose that at assessment k, we can afford to select m_k individuals from those who remain eligible after the earlier $k-1$ assessments. The second-order IPWEE for the responses upon the completion of assessment k is

$$U_{1k}(\bar{Y}_k|X;\bar{\theta}_k,\bar{\psi}_k^*) = \sum_{i=1}^{n} \frac{\bar{R}_{ik}^*}{\bar{\pi}_{ik}^*} \bar{D}_{i1k}' \bar{\Sigma}_{i1k}^{-1} (\bar{Z}_{i1k} - \bar{\xi}_{i1k}) = 0, \tag{5.11}$$

where

$$\bar{Z}_{i1k} = (Y_{i1}, \ldots, Y_{ik}, Y_{i1}Y_{i2}, \ldots, Y_{i(k-1)}Y_{ik})'$$

with $\bar{\xi}_{i1k} = E(\bar{Z}_{i1k}|X_i)$, $\bar{D}_{i1k} = \partial\bar{\xi}_{i1k}/\partial\bar{\theta}_k$ with $\bar{\theta}_k$ incorporating the marginal and association parameters β and $\bar{\alpha}_k$ induced by the longitudinal responses $\bar{Y}_k$, $\bar{\Sigma}_{i1k} = \mathrm{Cov}(\bar{Z}_{i1k}|X_i)$ is a working covariance matrix,

$$\bar{R}_{ik}^* = R_{i1} + (1 - R_{i1})R_{i2} + \ldots + (1 - R_{i1})(1 - R_{i2}) \ldots (1 - R_{i(k-1)})R_{ik},$$

and

$$\bar{\pi}_{ik}^* = \pi_{i1} + (1 - \pi_{i1})\pi_{i2} + \ldots + (1 - \pi_{i1})(1 - \pi_{i2}) \ldots (1 - \pi_{i(k-1)})\pi_{ik}$$

is indexed by $\bar{\psi}_k^*$ which is determined by ψ_k given $\bar{\psi}_{k-1}$. The estimating equation for the net selection model upon completion of the kth assessment is

$$U_{2k}(\bar{R}_k; \bar{\psi}_k^*) = \sum_{i=1}^{n} \frac{\partial \bar{\pi}_{ik}^*}{\partial \bar{\psi}_k^*} \frac{1}{\bar{\pi}_{ik}^*(1-\bar{\pi}_{ik}^*)} (\bar{R}_{ik}^* - \bar{\pi}_{ik}^*) = 0. \tag{5.12}$$

IPWEEs (5.11) and (5.12) allows for a prospective iteration as follows. Starting from $k = 1$, estimating equation $U_{11}(Y_1|X; \theta_1, \psi_1)$ with a conventional sampling scheme employed in the first assessment gives the preliminary parameter estimates. At assessment k, $k = 2, \ldots, K$, estimating equation $U_{1(k-1)}(\bar{Y}_{k-1}|X; \bar{\theta}_{k-1}, \bar{\psi}_{k-1}^*)$ updates the parameter estimates. The approximately optimal design for assessment k is then expected to minimize the asymptotic variance of estimator of β_1 derived from estimating equations $U_{1k}(\bar{Y}_k|X; \bar{\theta}_k, \bar{\psi}_k^*)$ and $U_{2k}(\bar{R}_k; \bar{\psi}_k^*)$ subject to the corresponding budgetary constraints. In other words, the optimal ψ_k is the one that minimizes the Lagrangian

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \bar{\vartheta}_k, \bar{\psi}_k] + \lambda \left[E(R_k|\bar{R}_{k-1} = 0; \bar{\vartheta}_k, \bar{\psi}_k) - \frac{E(M_k)}{E(N_k)} \right], \tag{5.13}$$

where $N_k = n - M_1 - \ldots - M_{k-1}$.

Time-Varying Exposures

If the exposure variable is time-dependent, then we can modify the response model as we did at the end of Section 5.3.1. Here we could consider the selection model

$$\pi_{ik} = P(R_{ik} = 1|\bar{R}_{i(k-1)}, \bar{Y}_{ik}, X_{i2}, \bar{X}_{i1(k-1)}^{\circ})$$

for assessment $k = 1, \ldots, K$, where $X_{i1j}^{\circ} = X_{i1j}$ if $R_{ij} = 1$ and carries no information otherwise, $j = 1, \ldots, k-1$. In other words, the marginal selection model at assessment k utilizes all the information up to assessment $k-1$, together with the response and auxiliary covariates at assessment k. Similarly, we could consider the pairwise log odds ratio to characterize the correlated selection indicators. The IPWEEs (5.8) and (5.10) can be adapted in a prospective way. Upon completion of the kth assessment, one could consider

$$U_{1k}(\bar{Y}_k|\bar{X}_k; \bar{\theta}_k, \bar{\psi}_k, \bar{\omega}_k) = \sum_{i=1}^{n} \bar{\Delta}_{ik} \bar{D}_{i1k}' \bar{\Sigma}_{i1k}^{-1} (\bar{Z}_{i1k} - \bar{\xi}_{i1k}) = 0 \tag{5.14}$$

with a weighting matrix

$$\bar{\Delta}_{ik} = \begin{pmatrix} \text{diag}\left(\frac{R_{ik}}{\pi_{ik}(\bar{Y}_{ik}, X_{i2}, \bar{X}^{\circ}_{i1(k-1)})}\right) & 0 \\ 0 & \text{diag}\left(\frac{R_{ij}R_{ik}}{\zeta_{ijk}(\bar{Y}_{ij}, \bar{Y}_{ik}, X_{i2}, \bar{X}^{\circ}_{i1(j-1)}, \bar{X}^{\circ}_{i1(k-1)})}\right) \end{pmatrix}.$$

Note that such a weighting matrix carries no information of selection beyond the kth assessment. In other words, it only involves the selection indicators $\bar{R}_k$ together with the corresponding marginal and association parameters of the selection models up to assessment k, $\bar{\psi}_k$ and $\bar{\omega}_k$. Furthermore, for the net selection model up to the kth assessment, one could consider

$$U_{2k}(\bar{R}_k; \bar{\psi}_k, \bar{\omega}_k) = \sum_{i=1}^{n} \bar{D}'_{i2k} \bar{\Sigma}^{-1}_{i2k} (\bar{Z}_{i2k} - \bar{\xi}_{i2k}) = 0, \tag{5.15}$$

where

$$\bar{Z}_{i2k} = (R_{i1}, \ldots, R_{ik}, R_{i1}R_{i2}, \ldots, R_{i(k-1)}R_{ik})'$$

with $\bar{\xi}_{i2k} = E(\bar{Z}_{i2k}|\bar{Y}_{ik}, X_{i2})$, $\bar{D}_{i2k} = \partial\bar{\xi}_{i2k}/\partial(\bar{\psi}'_k, \bar{\omega}'_k)$, and $\bar{\Sigma}_{i2k} = \text{Cov}(\bar{Z}_{i2k}|\bar{Y}_{ik}, X_{i2})$ is a working covariance matrix. The IPWEEs (5.14) and (5.15) allow for a similar prospective iteration as in the case of time-fixed exposures. As we move forward, each assessment of the longitudinal study involves an update of the preliminary parameter estimates followed by an approximately optimal sub-sampling scheme. Note that the Lagrangian for optimization now becomes

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \bar{\vartheta}_k, \bar{\psi}_k, \bar{\omega}_k] + \lambda \left[E(R_k|\bar{R}_{k-1} = 0; \bar{\vartheta}_k, \bar{\psi}_k, \bar{\omega}_k) - \frac{E(M_k)}{E(N_k)} \right].$$

References

Abu-Libdeh, H., Turnbull, B., and Clark, L. (1990). Analysis of multi-type recurrent events in longitudinal studies: application to a skin cancer prevention trial. *Biometrics*, 46:1017–1034.

Amorim, G. G. D. C. (2019). Semiparametric estimator for a secondary analysis with correlated outcomes. *Communications in Statistics-Theory and Methods*, 48(19):4703–4711.

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10(1):335–350.

Andersen, P. K., Klein, J. P., and Rosthoj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27.

Andersen, P. K. and Perme, M. P. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York: Wiley.

Bennett, P. and Burch, T. (1967). New York symposium on population studies in the rheumatic diseases: new diagnostic criteria. *Bulletin on Rheumatic Diseases*, 17:453––458.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forest. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*, volume 1. Taylor & Francis Group.

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

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48:457–468.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(8):429–36.

Cano, A., Baró, F., Fernández, C., Inaraja, V., and García-Domínguez, C. (2016). Evaluation of the risk factors of asymptomatic vertebral fractures in postmenopausal women with osteopenia at the femoral neck. *Maturitas*, 87:95–101.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., and Gladman, D. D. (2010a). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49:1399––1405.

Chandran, V., Tolusso, D. C., Cook, R. J., and Gladman, D. D. (2010b). Risk factors for axial inflammatory arthritis in patients with psoriatic arthritis. *The Journal of Rheumatology*, 37(4):809–815.

Chatterjee, N., Chen, Y.-H., and Breslow, N. (2003). A pseudo score estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98:158–168.

Chen, T. and Lumley, T. (2020). Optimal multiwave sampling for regression modeling in two-phase designs. *Statistics in Medicine*, 1:1–10.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters*, 49:155–161.

Cho, H., Jewell, N. P., and Kosorok, M. R. (2019). Interval censored recursive forests. *arXiv:1912.09983*.

Cook, R. and Lawless, J. (2019). Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics*, kxz047.

Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8(1):947–961.

Deville, J. C. and Sardnal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Efron, B. (1967). The two sample problem with censored data. *In Proc. 5th Berkeley Symposium on Math Statistics Probability Berkeley: University of California Press*, 4:831–853.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall.

Fu, W. and Simonoff, J. S. (2017). Survival trees for interval-censored survival data. *Statistics in Medicine*, 36(1):4831–4842.

Gaver, D. P. and Acar, M. a. (1979). Analytical hazard representations for use in reliability, mortality, and simulation studies. *Communication in Statistics. Simulation and Computation*, 8(2):91–111.

Ghosh, A., Wright, F. A., and Zou, F. (2013). Unified analysis of secondary traits in case–control association studies. *Journal of the American Statistical Association*, 108(502):566–576.

Gladman, D. D. and Chandran, V. (2011). Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology*, 50(1):25–31.

Gordon, L. and Olshen, R. (1985). Tree-structured survival analysis. *Cancer treatment report*, 69:1065–1069.

Han, S., Andrei, A. C., and Tsui, K. W. (2014). A semiparametric regression method for interval-censored data. *Communications in Statistics - Simulation and Computation*, 43(1):18–30.

Haneuse, S. and Rivera, C. (2018). On the analysis of case-control studies in cluster-correlated data settings. *Epidemiology*, 29(1):50–57.

Henrichon, E. and Fu, K. (1969). A nonparametric partitioning procedure for pattern classification. *IEEE Transactions on Computers*, C-18(7):614–624.

Hjorth, U. (1980). A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates. *Technometrics*, 22:99–107.

Hortobagyi, G. N., Theriault, R. L., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J., Knight, R. D., Heffernan, M., Reitsma, D. J., Kennedy, I., Allan, S. G., and Mellars, K. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases, protocol 19 aredia breast cancer study group. *New England Journal of Medicine*, 335:1785–1791.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Jiang, S. and Cook, R. J. (2020). Composite likelihood for aggregate data from clustered multistate processes under intermittent observation. *Communication in Statistics - Theory and Methods*, 49(12):2913–2930.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7:310–321.

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

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:413–438.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 48(2):411–425.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistics Association*, 88(422):457–467.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33:256–269.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Loh, W. (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348.

Lumley, T., Shao, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220.

McIssac, M. A. and Cook, R. J. (2013). Response-dependent sampling with clustered and longitudinal data. *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers, Lecture Notes in Statistics 211*, pages 157–181.

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

Meisel, W. and Michalopoulos, D. (1973). A partitioning algorithm with application in pattern classification and the optimization of decision trees. *IEEE Transactions on Computers*, C-22(1):93–103.

Molinaro, A. M., Dudoit, S., and van der Laan, M. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(422):154–177.

Neuhaus, J. M., Scott, A. J., Wild, C. J., Jiang, Y., McCulloch, C. E., and Boylan, R. (2014). Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics*, 70:44–52.

Neyman, J. (1938). Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association*, 33:101–116.

Palmgren, J. (1989). Regression models for bivariate binary responses. *UW Biostatistics Working Paper Series*, 101.

Pan, W. (1998). Rank invariant tests with left truncated and interval censored data. *Journal of Statistical Computation and Simulation*, 61:163–174.

Pan, Y., Cai, J., Kim, S., and Zhou, H. (2018a). Regression analysis for secondary response variable in a case-cohort study. *Biometrics*, 74:1014–1022.

Pan, Y., Cai, J., Longnecker, M. P., and Zhou, H. (2018b). Secondary outcome analysis for data from an outcome- dependent sampling design. *Statistics in Medicine*, 37:2321–2337.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 22:3035–3054.

Preisser, J. S. and Qaqish, B. F. (2014). A comparison of methods for simulating correlated binary variables with specified marginal means and correlations. *Journal of Statistical Computation and Simulation*, 84(11):2441–2452.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):825–839.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:18–84.

Raina, P., Wolfson, C., Kirkland, S., Griffith, L., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C., Hogan, D., Wister, A., Payette, H., Shannon, H., and Brazil, K. (2009). The Canadian longitudinal study on aging (CLSA). *Canadian Journal on Aging*, 28(3):221–229.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82:299–314.

Rivera-Rodriguez, C., Spiegelman, D., and Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38:4611—4624.

Robertson, T., Wright, F. T., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley: New York.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121.

Robins, J. M., Rotnitzky, A., Zhao, L. P., and Lipsitz, S. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rubin, D. and van der Laan, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3(1).

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

Saarela, O., Kulathinal, S., and Karvanen, J. (2012). Secondary analysis under cohort sampling designs using conditional likelihood. *Journal of Probability and Statistics*, 2012:1–37.

Schifano, E. D. (2019). A review of analysis methods for secondary outcomes in case-control studies. *Communications for Statistical Applications and Methods*, 26:103–129.

Schildcrout, J. S., Garbett, S. P., and Heagerty, P. J. (2013). Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 69(2):405–416.

Schildcrout, J. S., Haneuse, S., Tao, R., Zelnick, L. R., Schisterman, E. F., Garbett, S. P., Mercaldo, N. D., Rathouz, P. J., and Heagerty, P. J. (2019). Two-phase, generalized case-control designs for the study of quantitative longitudinal outcomes. *American Journal of Epidemiology. DOI: 10.1093/aje/kwz127*.

Schildcrout, J. S. and Heagerty, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9(4):735–749.

Schildcrout, J. S. and Heagerty, P. J. (2011). Outcome-dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis. *Biometrics*, 67(4):1583–1593.

Schildcrout, J. S., Schisterman, E. F., Mercaldo, N. D., Rathouz, P. J., and Heagerty, P. J. (2018). Extending the case-control design to longitudinal data: stratified sampling based on repeated binary outcomes. *Epidemiology*, 29(1):67–75.

Scott, A. and Wild, C. (1991). Fitting logistic regression models in stratified case–control studies. *Biometrics*, 1:497–510.

Scott, A. J. and Wild, C. J. (2011). Fitting regression models with response-biased samples. *The Canadian Journal of Statistics*, 39:519–536.

Steingrimsson, J. A., Diao, L., and Strawderman, R. L. (2019). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(1):370–383.

Strasser, H. and Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 8:220–250.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, volume 1. Springer Science+Business Media, Inc.

Tao, R., Mercaldo, N. D., Haneuse, S., Maronge, J. M., Rathouz, P. J., Heagerty, P. J., and Schildcrout, J. S. (2021). Two-wave two-phase outcome-dependent sampling designs, with applications to longitudinal binary data. *Statistics in Medicine*, 40:1863–1876.

Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case–control studies. *Biostatistics*, 15(1):117–128.

Therneau, T. and Atkinson, B. (2019). Recursive partitioning of classification, regression, and survival trees. *R package version 4.1-15*.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Science + Business Media, New York.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295.

Wellner, J. A. and Zhan, Y. (1997). A hybird algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, 92:945–959.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115:119–128.

Wu, Y. and Cook, R. J. (2020). Assessing the accuracy of predictive models with interval-censored data. *Biostatistics*, kxaa011.

Yao, W., Frydman, H., and Simonoff, J. S. (2019). An ensemble method for interval-censored time-to-event data. *Biostatistics*, kxz025.

Yin, Y. and Anderson, S. J. (2002). Nonparametric tree-structured modeling for interval-censored survival data. *In Proc. Biometrics Sections, Joint Statistical Meetings*.

Yu, Q., Li, L., and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, 27(1):35–44.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340.

Zhu, Y. Y., Lawless, J. F., and Cotton, C. A. (2017). Estimation of parametric failure time distributions based on interval-censored data with irregular dependent follow-up. *Statistics in Medicine*, 36:1548–1567.

APPENDICES

Appendix 2A: Additional Simulation Results for Chapter 2

Here present additional simulation results for an expansion of the simulation studies in Section 2.3. Various failure time distributions and additional underlying structures are considered to provide a fuller picture of the performance of our proposed methods.

We first adopt the covariates $X_1, \ldots, X_5$, the independent and highly correlated "autoregressive" dependence structures, and the data structure of the tree form with three terminal nodes in Section 2.3.1. Recall that the proportion of subjects falling into three terminal nodes are ensured to be 50%, 25%, and 25%, respectively. We also used the censoring mechanism discussed in Section 2.3.1 that addresses the heterogeneity of timings of assessments across subjects. Duration of follow-up, number of assessments, and assessment times are allowed to vary across individuals. Contrary to the settings shown in Table 2.1, we now consider failure times under the terminal nodes that follow Weibull distributions with decreasing hazards, log-normal distributions, and distributions with a bathtub-shaped hazard; see Table 2A1 for a summary. The parameter configuration of the Weibull and log-normal distributions come from the simulation studies reported in Fu and Simonoff (2017). The two-parameter distribution with a bathtub-shaped hazard function is proposed by Chen (2000), where the cumulative distribution function is

$$F(t) = 1 - \exp\{\lambda[1 - \exp(t^{\beta})]\},$$

and the hazard function

$$h(t) = \lambda\beta t^{\beta-1}\exp(t^{\beta})$$

has a bathtub shape when $\beta < 1$. Such distributions with a bathtub-shaped hazard are commonly used in reliability analysis. See Hjorth (1980) and Gaver and Acar (1979) for alternative parametric models with bathtub-shaped hazards. We consider a sample size $n = 200$ with 500 replications. The aim is to compare the performance in predicting failure times, recovering the underlying structure, and predicting the failure status of our proposed regression trees based on CUT and PO imputations with those of the oracle regression tree, the regression trees following from midpoint and right endpoint imputations, and the conditional inference tree.

Table 2A1: Failure time distributions under the terminal nodes. The brackets display the parameters of the distributions.

Failure time distributions	Node 1	Node 2	Node 3
Weibull (κ, λ)	(0.9, 1)	(0.9, 3)	(0.9, 7)
Log-normal (μ, σ)	(0.5, 0.5)	(1.3, 0.3)	(2, 0.3)
Bathtub (β, λ)	(0.8, 2)	(0.8, 0.5)	(0.5, 0.1)

The performances in predicting failure times and failure status at 0.25, 0.50, and 0.75 quantiles of the marginal distribution of the failure times T are shown in Figures 2A1, 2A2, and 2A3. Following the order of the figures, the failure times follow Weibull distributions with a decreasing hazard, log-normal distributions, and distributions with a bathtub-shaped hazard. Within each figure, two columns from left to right represent settings with independent and highly correlated covariates, respectively. Four rows from top to bottom display results of predicting failure times and failure status at the three quantiles, respectively. The evaluation metrics are the prediction errors, PE (Section 2.3.2) for failure times and $PE_{survivor}$ (Section 2.3.3) for failure status at a fixed time horizon. In each subfigure, the order of the boxplots follows from Section 2.3, and the results of the right endpoint imputation are not shown because of the dramatically large PEs. Our proposed CUT and PO imputations are comparable to the oracle trees when predicting failure

times in all three settings. The midpoint imputation also predicts failure times comparably well to the oracle trees. The conditional inference trees give large prediction errors. As for failure status, all methods are comparable other than the traditional imputations when the failure times follow Weibull distributions with decreasing hazards or distributions with bathtub-shaped hazards. The conditional inference trees have the lowest prediction error in the third row of Figure 2A2 when the covariates are independent and in the third row of Figure 2A3. However, as shown in the last row of both Figure 2A2 and Figure 2A3, they produce large prediction errors at the 0.75 quantile.

We display the performance in structure recovery in Table 2A2. The evaluation metrics are "model size", " number of predictors", "percent correct", and "percent without noise" as explained in Section 2.3.2. The setting in which failure times follow Weibull distributions with decreasing hazards lead to worse performance than the other two settings in general. Our proposed regression trees recover the underlying structure and avoid picking up noise variables comparably well to the oracle trees in all three settings. The traditional imputations perform comparably well to the oracle trees throughout the settings, too. The conditional inference trees recover the true structure less frequently, except in the log-normal setting with independent covariates.

Next, we expand the true relationship between the failure times and covariates beyond a tree structure to see how well the nonparametric nature of the tree adapts to simpler and more complex structures. Given the multivariate normals $(W_1, \ldots, W_5)$ generated as in Section 2.3.1, we consider failure times that follow Weibull distributions whose shape parameters are fixed, and scale parameters are functions of covariates $X_1 = I(W_1 < 0)$ and $X_4 = e^{-|W_4|}$. Covariates X_2, X_3, and X_5 specified in Section 2.3.1 are still involved in the training and test datasets as noise variables. Specifically, we let

- $\theta_1 = -X_1 - X_4$;
- $\theta_2 = -\cos[(X_1 + X_4)\pi] - \sqrt{X_1 + X_4}$,

and for $\theta \in \{\theta_1, \theta_2\}$, we consider failure times that follow

- Weibull distribution with an increasing hazard, $\text{Wei}(2, 10e^{\theta})$;

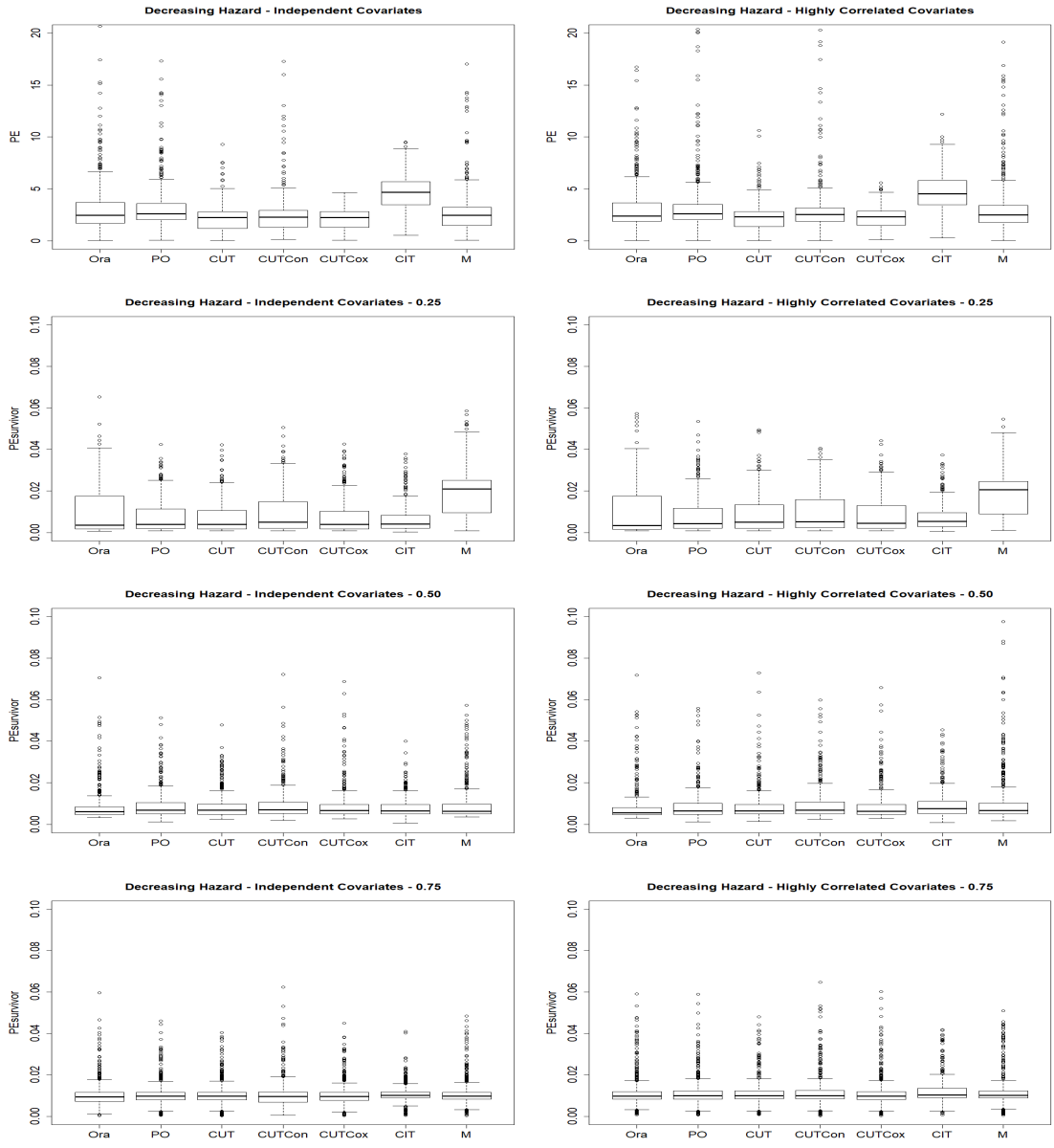

Figure 2A1: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M). The Weibull failure times under the terminal nodes have decreasing hazards.

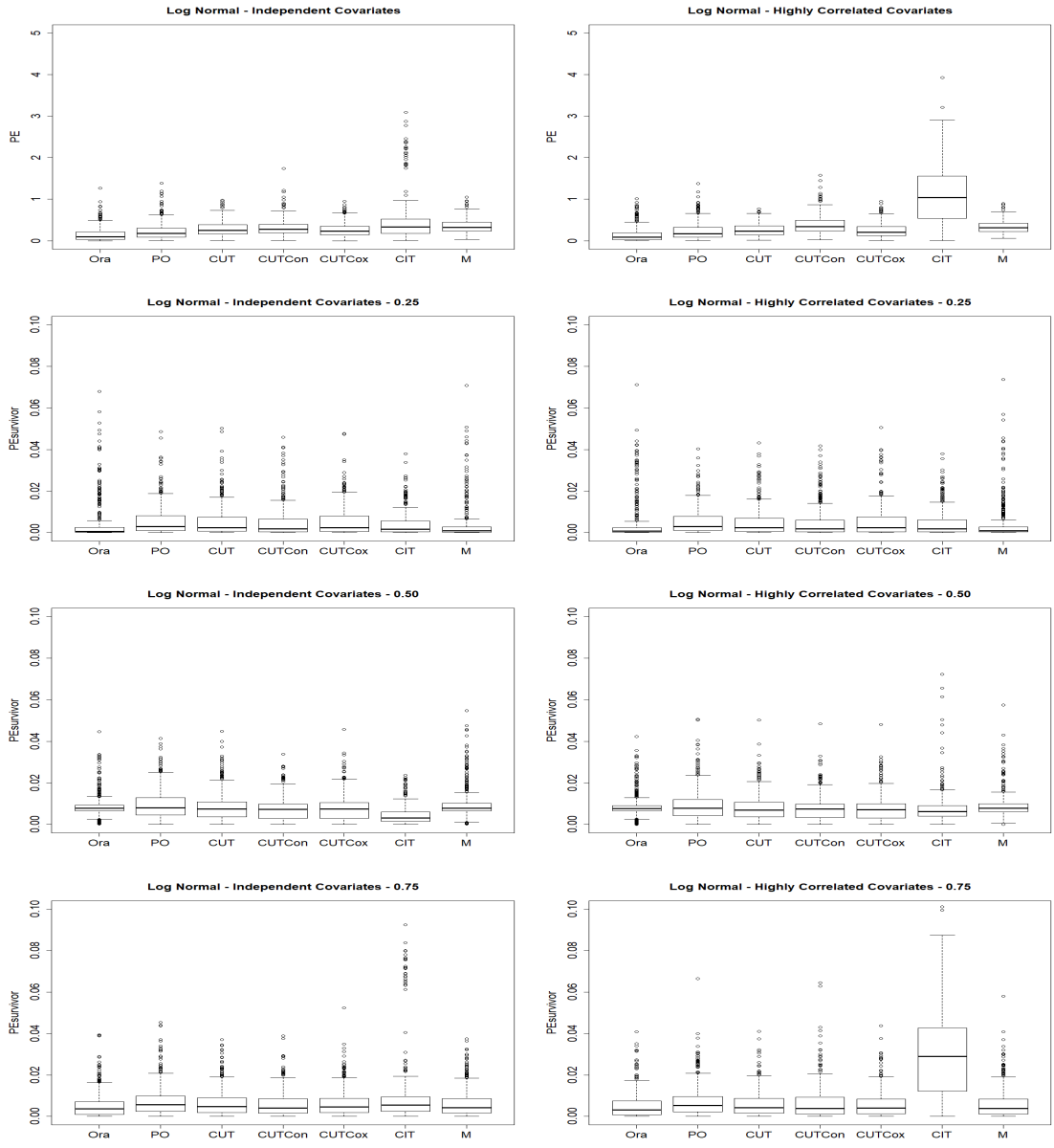

Figure 2A2: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M). The failure times under the terminal nodes follow log-normal distributions.

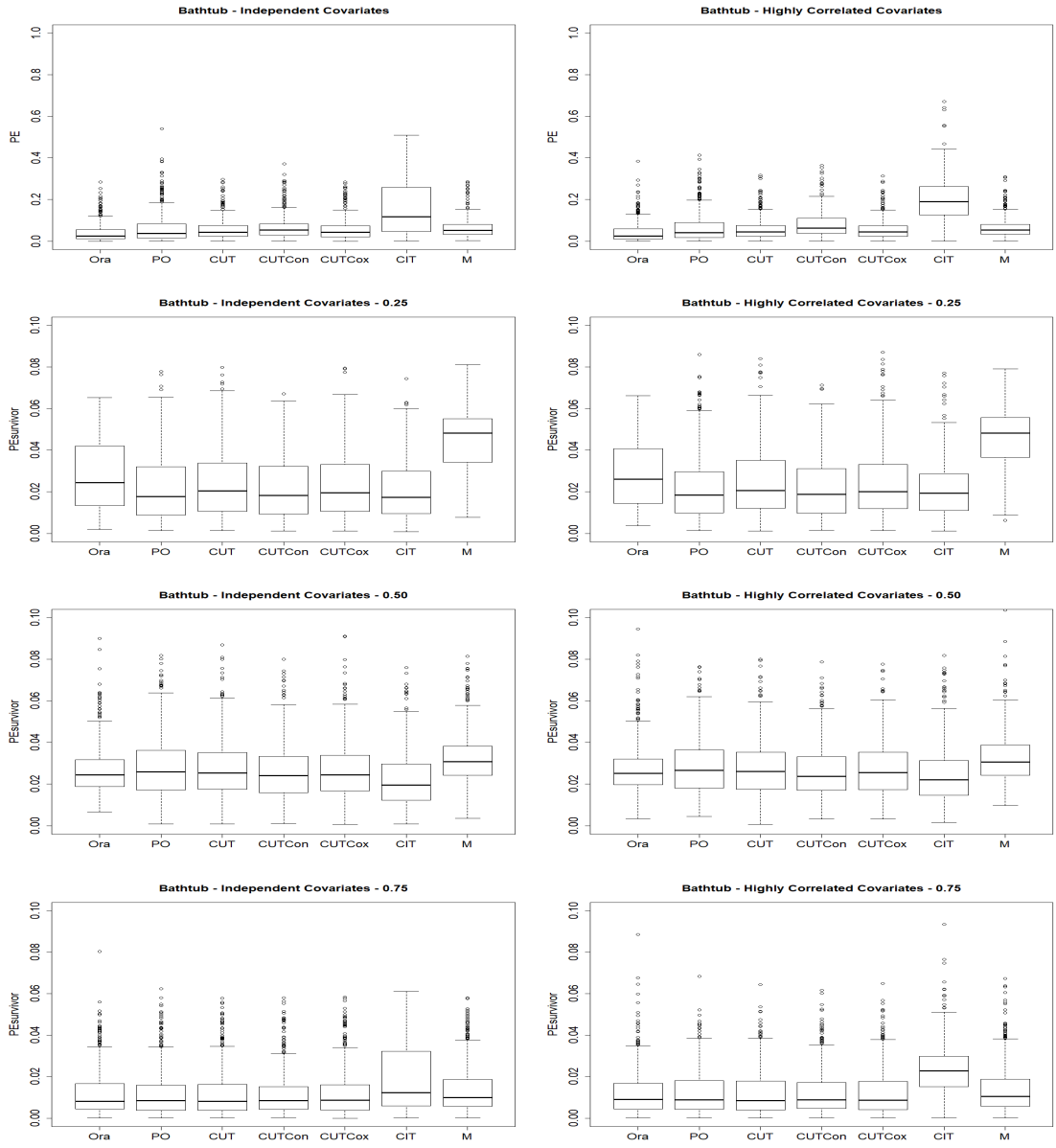

Figure 2A3: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M). The failure times under the terminal nodes have bathtub-shaped hazards.

Table 2A2: Structure recovery measures with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Oracle, CIT, M, R). Failure time distributions under the terminal nodes follow Weibull distribution with decreasing hazards or log-normal distributions or distributions with bathtub-shaped hazards.

	Independent Covariates								Highly Correlated Covariates							
	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R
Weibull distributions with decreasing hazards																
Model Size	3.320	2.878	2.836	2.898	2.890	2.304	2.694	2.736	3.278	2.844	2.840	2.956	2.994	2.490	2.832	2.754
# Predictors	1.938	1.682	1.684	1.728	1.728	1.284	1.632	1.618	1.888	1.614	1.632	1.684	1.696	1.446	1.610	1.600
% Correct	45.4	39.8	44.6	43.8	44.2	16.8	45.4	40.2	38.6	32.4	36.0	30.8	36.0	10.0	37.6	33.2
% w/o Noise	77.0	87.0	89.0	86.8	86.4	92.8	91.8	90.0	72.4	81.6	83.4	74.8	80.2	65.4	85.2	82.8
Log-normal distributions																
Model Size	3.242	3.174	3.280	3.334	3.362	3.080	3.216	3.220	3.216	3.280	3.256	3.545	3.404	3.758	3.194	3.204
# Predictors	2.112	2.084	2.104	2.122	2.122	2.016	2.076	2.076	2.092	2.106	2.078	2.265	2.138	2.633	2.062	2.080
% Correct	90.4	93.0	91.6	90.4	90.0	89.8	93.6	94.0	91.8	90.4	92.6	75.4	88.4	22.6	94.0	92.8
% w/o Noise	90.4	93.0	91.6	90.4	90.0	94.0	93.6	94.0	91.8	90.4	92.6	75.4	88.4	22.6	94.0	92.8
Bathtub-shaped hazards																
Model Size	3.338	3.278	3.246	3.307	3.304	2.814	3.190	3.190	3.342	3.432	3.432	3.520	3.488	3.337	3.366	3.304
# Predictors	2.126	2.104	2.108	2.118	2.124	1.782	2.082	2.078	2.120	2.132	2.140	2.197	2.194	2.245	2.124	2.074
% Correct	88.2	85.4	88.0	86.0	86.8	64.4	90.2	89.4	87.8	81.4	83.8	74.8	80.8	22.8	86.0	86.4
% w/o Noise	88.6	88.4	89.4	88.4	88.6	92.6	91.6	91.4	88.0	83.4	84.8	76.0	81.6	27.2	87.0	88.6

- Weibull distribution with a decreasing hazard, Wei$(0.5, 5e^{\theta})$.

The settings are similar to those in Fu and Simonoff (2017). Note that the linear proportional hazards assumption is satisfied in the first setup with $\theta = \theta_1$. The second setting where $\theta = \theta_2$ aims to test the effectiveness of our proposed methods in a real-world application in which failure times may have a complex structure. For convenience, we denote the two settings as the proportional hazard setting and the complex setting, respectively.

Adopting the censoring mechanism discussed in Section 2.3.1, we considered a sample size $n = 200$ with 500 replications. The prediction performance of failure times and failure status at 0.25, 0.50, and 0.75 quantiles of the marginal distribution of T are assessed by prediction errors and reported in Figures 2A4, 2A5, 2A6, and 2A7. The former two show results of the proportional hazard setting. The Weibull distributions have increasing hazards in Figure 2A4 and decreasing hazards in Figure 2A5. Similarly, the latter two show results of the complex setting. The Weibull distributions have increasing hazards in Figure 2A6 and decreasing hazards in Figure 2A7. In the proportional hazard setting, the CUT methods based on the conditional inference tree and Cox model estimations of the conditional survivor function are comparable to the oracle tree. The midpoint imputation, PO imputation, and the CUT method employing the marginal survivor function give larger prediction errors, especially when the Weibull failure distributions have increasing hazards. The conditional inference tree predicts failure status comparably well to the oracle trees regardless of the dependence structures of the covariates. However, it does not do well when predicting failure times. In the complex setting, the CUT method based on the conditional inference tree and Cox model estimations of the conditional survivor function are the only two approaches comparable to the oracle tree when the Weibull failure distributions have increasing hazards. As for the decreasing hazards case, all methods are comparable to the oracle trees, except that the conditional inference tree does not perform well in predicting failure times, and the midpoint imputation does not do well in predicting failure status.

Since the underlying structure is not of a tree form, the evaluation metrics "model size" for structure recovery is not applicable here. However, we report "number of predictors", "percent correct" and "percent without noise" in Table 2A3 for completeness. As θ is a function of X_1 and X_4, we define fitted trees that split on both the covariates and nothing

else as "correct", regardless of how many times they appear in the tree models. Similarly, fitted trees that do not split on X_2, X_3, and X_5 are considered as "without noise". All the methods are very conservative and can hardly catch both the influential predictors when the Weibull failure time distributions have decreasing hazards in both the proportional hazard and complex settings, although they avoid picking up noise variables most of the times. When the failure time distributions have increasing hazards, our proposed CUT methods are comparable to the oracle trees in both the proportional hazard and complex settings. The PO imputations and the midpoint imputations pick up both influential predictors less frequently when the covariates are highly correlated, while the right endpoint imputation is too conservative to do so regardless of the dependence structure of the covariates. Still, the PO and traditional imputations avoid picking up noise variables comparably to the oracle trees in most settings. Finally, the conditional inference trees catch the influential predictors most frequently when the Weibull failure times have increasing hazards in the proportional hazard setting. Furthermore, we note that the dependence structure of the covariates does not affect the performance of the conditional inference trees as it does when the underlying structure has a tree form.

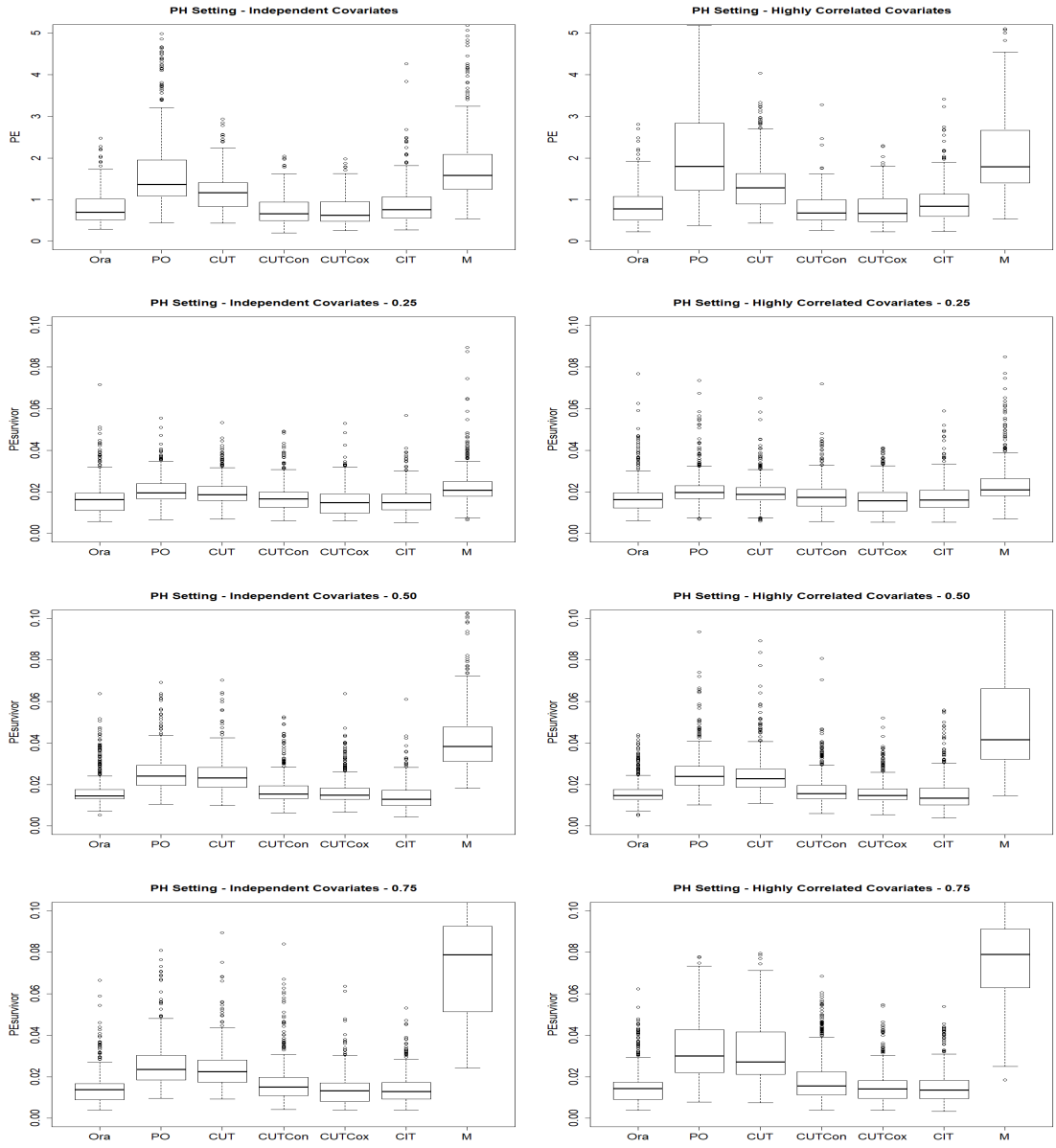

Figure 2A4: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M) in the proportional hazard setting. The Weibull failure times have increasing hazards.

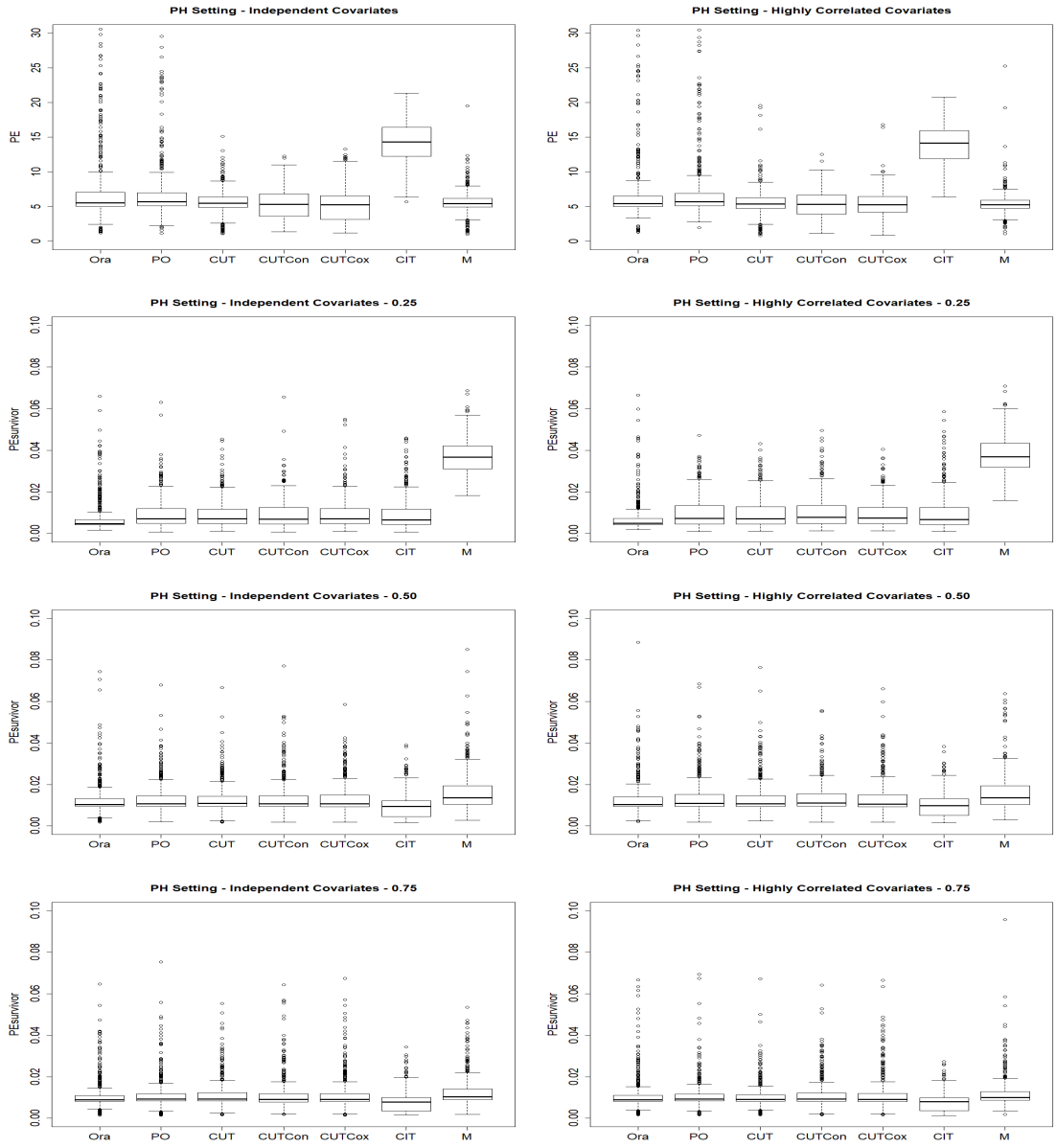

Figure 2A5: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M) in the proportional hazard setting. The Weibull failure times have decreasing hazards.

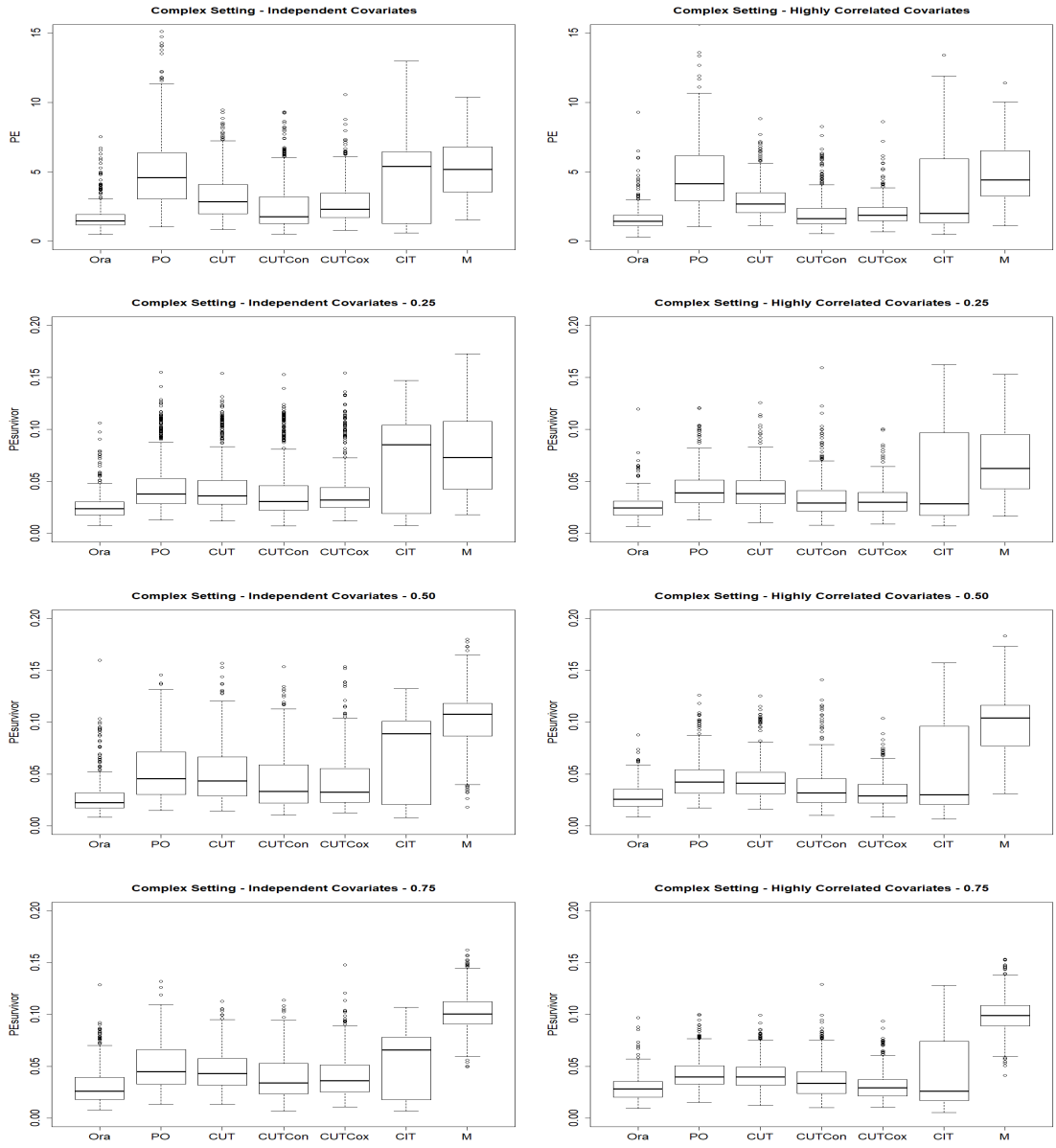

Figure 2A6: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M) in the complex setting. The Weibull failure times have increasing hazards.

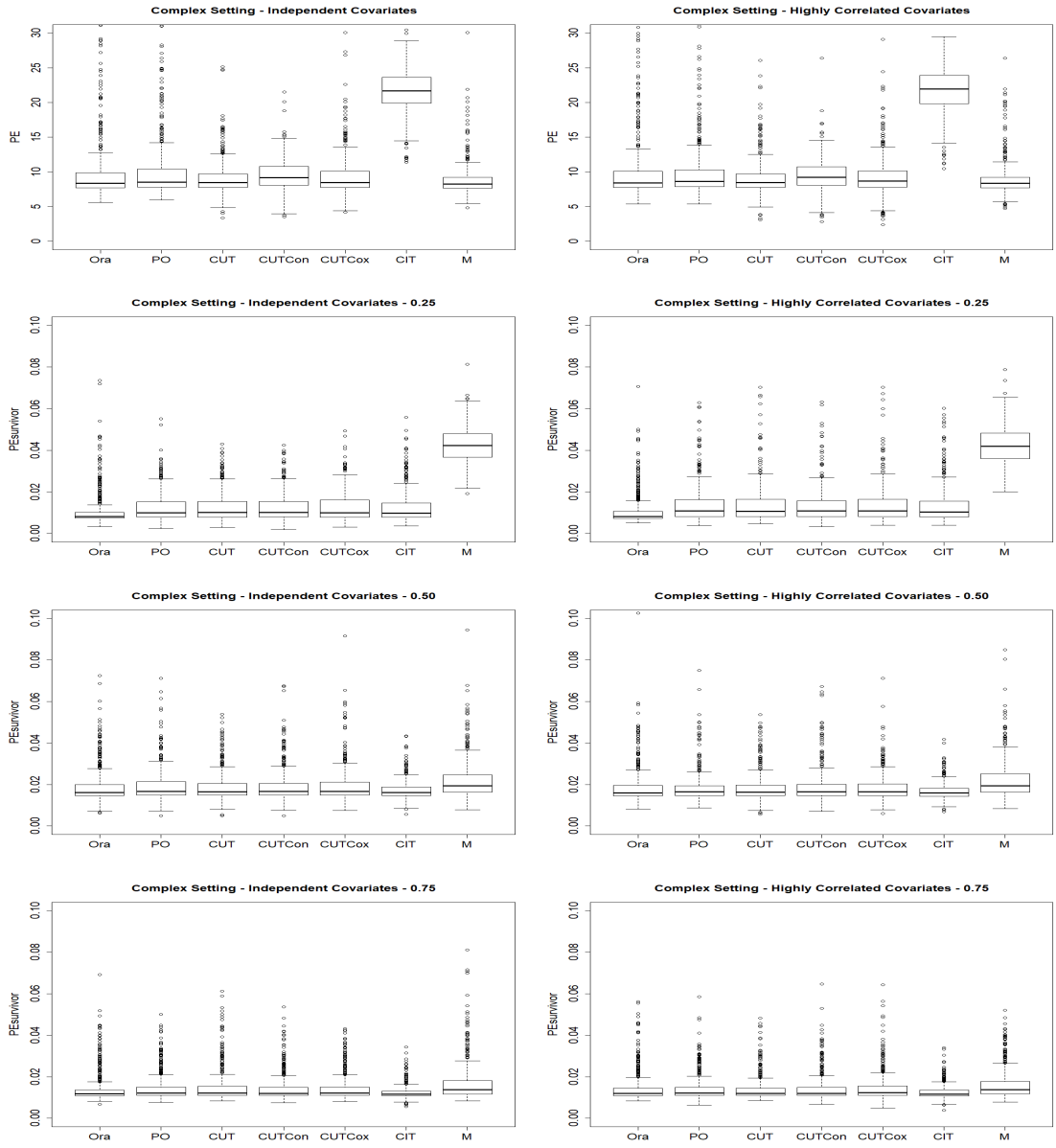

Figure 2A7: Prediction errors for predicting failure times and failure status with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Ora, CIT, M) in the complex setting. The Weibull failure times have decreasing hazards.

Table 2A3: Structure recovery measures with independent and highly correlated covariates comparing proposed CART algorithms (PO, CUT, CUT_{Con}, CUT_{Cox}) for interval-censored failure time data and the benchmarks (Oracle, CIT, M, R). Failure times follow Weibull distributions with fixed shape parameters and scale parameters as functions of covariates.

	Independent Covariates								Highly Correlated Covariates							
	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R	Oracle	PO	CUT	CUT_{con}	CUT_{cox}	CIT	M	R
Proportional hazard setting with increasing hazards																
# Predictors	2.088	1.986	1.836	2.098	2.148	2.088	1.776	0.502	2.140	2.070	2.090	2.195	2.232	2.337	1.830	0.558
% Correct	59.6	48.4	56.2	63.8	59.8	85.2	53.6	4.8	48.6	20.8	36.4	53.0	51.8	59.8	26.2	2.0
% w/o Noise	79.2	77.6	87.2	80.0	77.4	88.2	88.2	90.6	66.4	40.2	60.0	67.6	66.6	62.6	51.6	69.2
Proportional hazard setting with decreasing hazards																
# Predictors	0.500	0.362	0.476	0.906	0.846	0.772	0.388	0.300	0.378	0.334	0.446	0.871	0.696	0.777	0.350	0.310
% Correct	3.0	2.4	4.4	7.6	7.4	5.6	1.2	0.0	1.0	1.0	1.6	2.2	2.6	1.8	0.6	0.2
% w/o Noise	86.2	89.0	90.4	88.8	84.6	92.4	87.6	87.4	81.2	82.8	77.4	72.2	73.2	77.6	80.4	84.4
Complex setting with increasing hazards																
# Predictors	2.780	2.028	2.316	2.445	2.474	1.084	1.600	0.448	2.868	2.484	2.788	2.766	2.936	1.796	2.120	0.460
% Correct	47.0	43.2	56.4	55.8	60.0	43.2	44.6	1.6	30.4	22.8	28.2	32.4	31.2	37.0	22.4	0.6
% w/o Noise	47.4	66.8	64.8	61.8	63.0	92.2	79.0	88.6	30.6	36.6	30.0	33.8	31.4	64.4	46.0	86.8
Complex setting with decreasing hazards																
# Predictors	0.376	0.274	0.352	0.596	0.512	0.260	0.268	0.318	0.400	0.360	0.530	0.640	0.706	0.340	0.316	0.290
% Correct	2.8	1.8	3.8	8.6	3.4	6.2	1.4	0.4	2.6	2.2	4.4	6.6	5.2	7.0	1.2	0.2
% w/o Noise	87.8	91.2	87.8	84.6	83.2	95.0	89.6	86.2	86.4	87.6	79.2	80.4	76.6	91.8	87.2	86.8

Appendix 3A: Multiple Stages of Phase II Sub-Sampling in Adaptive Two-Phase Designs

The adaptive two-phase designs discussed can be extended to accommodate multiple interim stages of phase II sub-sampling labelled as, say, phase IIA, IIB, IIC, and so on. We next describe how this can be implemented for the maximum likelihood and IPWEE approaches when we add one more phase to the phase II process. Let $\bar{\vartheta}_C = (\bar{\beta}_C', \bar{\gamma}_C')'$ denote the parameter estimates following phases IIA and IIB which are used to find the approximately optimal ψ_C for the selection model of a phase IIC sub-sample of $M_C = M - M_A - M_B$ individuals. Following the maximum likelihood approach the score vector for ϑ_1 becomes

$$S_C(Y, X_1|X_2; \vartheta_1) = \sum_{i=1}^{n} S_{iC}(Y_i, X_{i1}|X_{i2}; \vartheta_1) = \sum_{i=1}^{n} (S_{i1C}'(\vartheta_1), S_{i2C}'(\vartheta_1))',$$

where

$$\begin{aligned} S_{i1C}(\vartheta_1) &= R_i^C \mathcal{S}_{i1}(Y_i|X_i;\beta) + (1-R_i^C) E_{X_{i1}|Y_i,X_{i2}}(\mathcal{S}_{i1}(Y_i|X_i;\beta);\vartheta_1) \\ S_{i2C}(\vartheta_1) &= R_i^C \mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1) + (1-R_i^C) E_{X_{i1}|Y_i,X_{i2}}(\mathcal{S}_{i2}(X_{i1}|X_{i2};\gamma_1);\vartheta_1)\ , \end{aligned}$$

with

$$R_i^C = A_i + (1-A_i)B_i + (1-A_i)(1-B_i)C_i,$$

for $i = 1, \ldots, n$. Under regularity conditions, the solution to $S_C(Y, X_1|X_2; \vartheta_1) = 0$ is asymptotically normal and

$$\sqrt{n}(\hat{\vartheta}_1 - \vartheta_1) \to N(0, E[S_{iC}(Y_i, X_{i1}|X_{i2};\vartheta_1) S_{iC}'(Y_i, X_{i1}|X_{i2};\vartheta_1)]^{-1}),\ \text{as } n \to \infty,$$

with expectations taken with respect to Y, X_1, X_2, A, B, and C. The approximately optimal ψ_C is defined as the one that minimizes

$$\text{asvar}[\sqrt{n}(\hat{\beta}_1 - \beta_1); \bar{\vartheta}_C, \psi_A, \psi_B, \psi_C] + \lambda \left[E(C|A=B=0; \bar{\vartheta}_C, \psi_A, \psi_B, \psi_C) - \frac{E(M_C)}{E(N_C)} \right], \quad (16)$$

where $N_C = n - M_A - M_B$.

Similarly, the IPWEEs become

$$\begin{aligned}\sum_{i=1}^{n} U_{i1C}(Y_i|X_i;\beta,\psi^{C*}) &= \sum_{i=1}^{n} \frac{R_i^C}{\pi_i^{C*}} D_{i1}' \Sigma_{i1}^{-1}(Y_i-\mu_i) = 0 \\ \sum_{i=1}^{n} U_{i2C}(A_i,B_i,C_i;\psi^{C*}) &= \sum_{i=1}^{n} \frac{\partial \pi_i^{C*}}{\partial \psi^{C*}} \frac{1}{\pi_i^{C*}(1-\pi_i^{C*})}(R_i^C-\pi_i^{C*}) = 0,\end{aligned}$$

where

$$\pi_i^{C*} = \pi_{iA} + (1-\pi_{iA})\pi_{iB} + (1-\pi_{iA})(1-\pi_{iB})\pi_{iC}$$

is indexed by ψ^{C*} which is determined by ψ_A and ψ_B in phases IIA and IIB, as well as ψ_C to be set in phase IIC. The optimization problem is now to minimize (16) in which the asymptotic covariance matrix has the sandwich form. The generalization to accommodate more interim stages are straightforward in principle but more computationally intensive.

We can assess the efficiency gains or losses for the maximum likelihood and IPWEE approaches as a function of the number of interim stages by generalizing the simulations for binary X_1 to allow for two to four interim stages of phase II sub-sampling. We compare several simulation set-ups with $n = 5000$, $E(M) = 2000$, and $E(M_A) = 500$ having two to four stages in phase II. Specifically, the setting with $E(M_B) = 1500$ is compared to those with $E(M_A) = E(M_B) = 500$, $E(M_C) = 1000$ and $E(M_A) = E(M_B) = E(M_C) = E(M_D) = 500$. According to Table 3A1, having additional stages in phase II give slightly better results in some cases. The comparison among set-ups with $E(M_A) : E(M_B) : E(M_C) : E(M_D)$ = 1:3:0:0, 1:1:2:0, and 1:1:1:1 was repeated for smaller phase II sub-samples with $E(M) = 1000$ and $E(M) = 400$. The efficiency gains in all settings were negligible and so it is difficult to justify having multiple interim stages in phase II when the phase II sub-sample size is modest.

Table 3A1: Average standard errors (ASE), empirical standard errors (ESE), and percent empirical coverage probabilities (ECP%) of estimators from maximum likelihood (ML) and IPWEE (IPW) adaptive two-phase designs with SRS or BS employed in a phase IIA sub-sample of size $0.25E(M)$. $E(M_A) : E(M_B) : E(M_C) : E(M_D)$ = 1:3:0:0, 1:1:2:0, and 1:1:1:1 from left to right. Phase I sample size $n = 5000$, phase II sub-sample size $E(M) = 2000$, and $nsim = 1000$. Parameter of interest $\beta_1 = 0.916$.

Phase IIA		No. of stages in phase II								
		2			3			4		
% Sampling*	Analysis	ASE	ESE	ECP%	ASE	ESE	ECP%	ASE	ESE	ECP%
25 SRS	ML	0.108	0.108	94.7	0.108	0.107	96.1	0.107	0.104	95.9
	IPW	0.110	0.114	94.4	0.110	0.111	94.8	0.110	0.111	95.5
25 BS	ML	0.108	0.107	94.5	0.108	0.109	94.3	0.108	0.109	94.3
	IPW	0.111	0.114	94.7	0.110	0.112	94.7	0.110	0.110	95.5

* Percentage of the phase II sub-sample chosen from and the sampling scheme employed in phase IIA.

Appendix 4A: Additional Simulation Results for Chapter 4

Here we present simulation results of the sequential two-phase designs discussed in Section 4.3 in a setting where the responses are assumed to have an exchangeable odds ratio. In other words, we assume

$$\log \phi(X, Z) = \gamma_0$$

in data generation, design, and analysis. We adopt the parameter configuration and the size of the empirical studies presented in Section 4.3.2. We specify $\gamma_0 = \log 2$ and $\gamma_0 = \log 4$ to reflect a moderate and a strong association between the responses, respectively. Table 4A1 summarizes the empirical biases (EBias), asymptotic standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of designs **A-D**. Figure 4A1 plots the asymptotic standard error of $\hat{\beta}_1$ of designs $\mathbf{D_1}$ as a function of the proportion of the individuals in the combined phase II sub-sample that are selected from Study 1. Table 4A2 summarizes the sampling probabilities of strata (Y, Z) of designs $\mathbf{D_2}$ and $\mathbf{D_1}$ displayed in Table 4A1.

Table 4A1: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the estimator of parameter of interest from the combined phase II data using likelihood and inverse weighting methods. **A-D** refer to designs with different use of Study 1 data described in Section 4.3.2. For designs **A**-$\mathbf{D_1}$, $E(M_1) = E(M_2) = 0.05n$. For designs $\mathbf{D_2}$, $E(M_2) = 0.1n$. "BS" and "opt" stand for balanced sampling and optimal sampling, respectively. An exchangeable dependence structure is employed in data generation, design, and analysis. Moderate and strong associations between the responses are reflected by $\phi(X, Z) = 2$ and $\phi(X, Z) = 4$, respectively. $nsim = 1000$, $n = 5000$, and $\beta_1 = 0.916$.

	Stratification				Results			
Design	Study 1	Study 2	Response Model	Analysis	EBias	ASE	ESE	ECP(%)
			$\phi(X, Z) = 2$					
A	-	BS (Y_2, Z)	Marginal	IPW	0.035	0.351	0.354	95.7
				ML	0.018	0.298	0.297	95.9
				CML	0.018	0.299	0.297	95.9
B	BS (Y_1, Z)	BS (Y_2, Z)	Marginal	IPW	-0.062	0.239	0.245	92.9
				ML	-0.043	0.207	0.213	93.2
				CML	-0.043	0.206	0.213	93.2
C	BS (Y_1, Z)	BS (Y, Z)	Joint	IPW	0.009	0.271	0.269	94.8
				IPW2	0.008	0.245	0.241	95.0
				ML	0.004	0.203	0.200	95.7
				CML	0.004	0.205	0.203	95.3
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	Joint	IPW	0.010	0.227	0.222	96.0
				IPW2	0.007	0.220	0.215	96.0
				ML	0.005	0.194	0.195	94.8
				CML	0.003	0.196	0.196	95.2
$\mathbf{D_2}$	-	opt (Y_2, Z)	Marginal	IPW	-	0.221	-	-
				ML	-	0.184	-	-
				CML	-	0.184	-	-
			$\phi(X, Z) = 4$					
A	-	BS (Y_2, Z)	Marginal	IPW	0.048	0.351	0.364	95.0
				ML	0.026	0.299	0.309	94.8
				CML	0.026	0.298	0.309	94.7
B	BS (Y_1, Z)	BS (Y_2, Z)	Marginal	IPW	-0.049	0.238	0.235	95.0
				ML	-0.034	0.206	0.207	94.3
				CML	-0.034	0.206	0.207	94.3
C	BS (Y_1, Z)	BS (Y, Z)	Joint	IPW	-0.006	0.266	0.271	94.2
				IPW2	-0.007	0.243	0.248	94.6
				ML	-0.004	0.205	0.213	94.0
				CML	-0.003	0.207	0.215	93.5
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	Joint	IPW	0.009	0.227	0.222	95.5
				IPW2	0.008	0.220	0.217	94.5
				ML	-0.001	0.194	0.200	94.2
				CML	-0.001	0.196	0.201	94.9
$\mathbf{D_2}$	-	opt (Y_2, Z)	Marginal	IPW	-	0.221	-	-
				ML	-	0.184	-	-
				CML	-	0.184	-	-

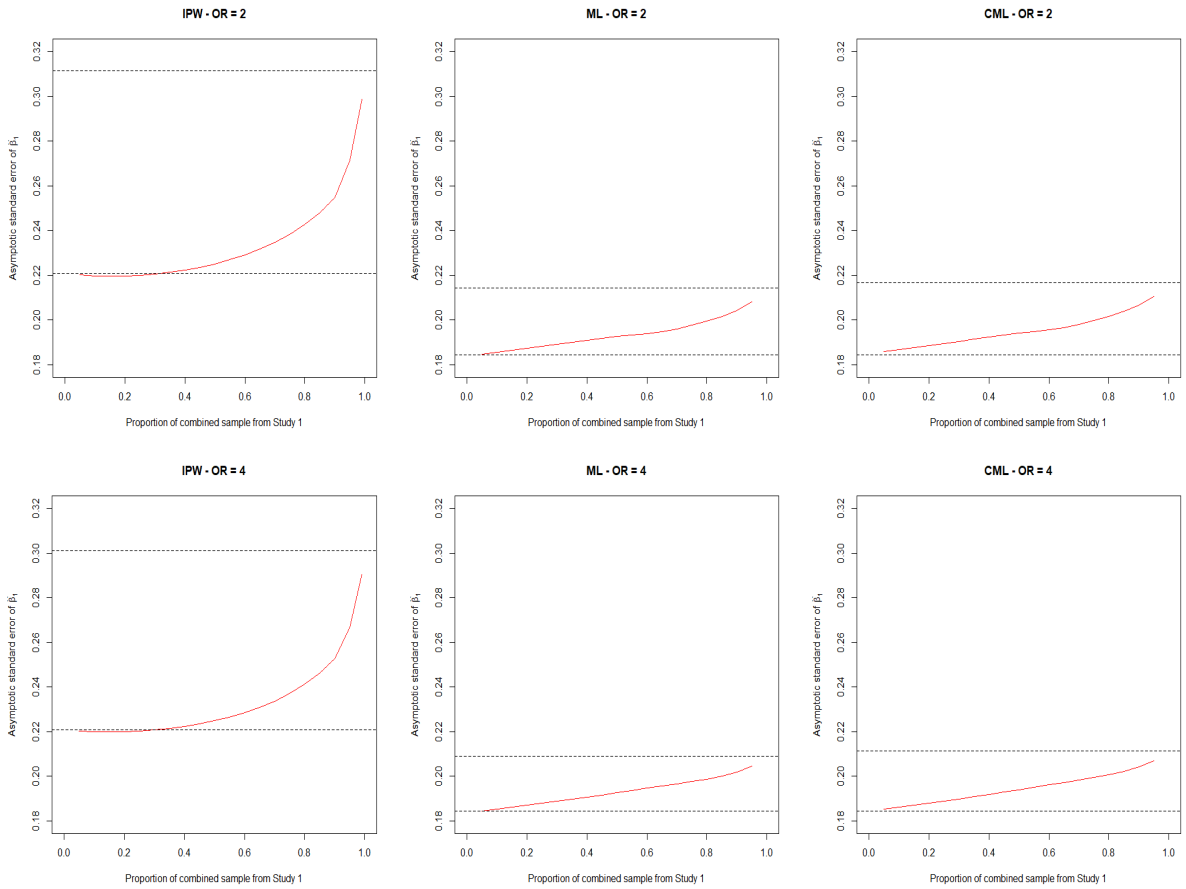

Figure 4A1: Plots of asymptotic standard error of $\hat{\beta}_1$ of designs $\mathbf{D_1}$ as the proportion of the individuals in the combined phase II sub-sample that are selected from Study 1, $E(M_1)/E(M_1+M_2)$, increases. An exchangeable dependence structure is employed in data generation, design, and analysis. The two rows display graphs of the set-ups with odds ratio (OR) 2 and 4, respectively. Columns from left to right display graphs of frameworks of analysis IPW, ML, and CML, respectively. Lower bounds represent the ideal designs $\mathbf{D_2}$ which select an optimal sample of $M_1 + M_2$ individuals in Study 2. Upper bounds represent using BS to select $M_1 + M_2$ individuals in Study 1. $nsim = 1000$, $n = 5000$, $E(M_1 + M_2) = 500$.

Table 4A2: Sampling probabilities of 8 strata defined by (Y_1, Y_2, Z) of our proposed optimal designs $\mathbf{D_1}$ and the ideal optimal designs $\mathbf{D_2}$. The Study 1, Study 2, and Net Study 2 rows refer to π_1, π_2, and $\bar{\pi}_2$ of the proposed designs $\mathbf{D_1}$, respectively. An exchangeable dependence structure is employed in data generation, design, and analysis. Moderate and strong associations between the responses are reflected by $\phi(X, Z) = 2$ and $\phi(X, Z) = 4$, respectively. $nsim = 1000$, $n = 5000$, and $\beta_1 = 0.916$.

Analysis	Designs	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,0)	(1,1,1)
					$\phi(X,Z)=2$				
Expected Strata Size		2788	504	504	539	169	169	204	123
					Selection Probabilities				
IPW	$\mathbf{D_2}$	0.063	0.063	0.206	0.092	0.092	0.281	0.206	0.281
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.033	0.032	0.169	0.016	0.016	0.089	0.167	0.087
	$\mathbf{D_1}$ Net Study 2	0.051	0.118	0.185	0.103	0.226	0.169	0.240	0.282
ML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.011	0.011	0.050	0.371	0.370	0.114	0.872
	$\mathbf{D_1}$ Net Study 2	0.020	0.099	0.030	0.134	0.505	0.426	0.192	0.899
CML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.011	0.011	0.050	0.374	0.373	0.108	0.876
	$\mathbf{D_1}$ Net Study 2	0.020	0.098	0.029	0.134	0.508	0.429	0.187	0.902
					$\phi(X,Z)=4$				
Expected Strata Size		2863	429	429	569	139	139	279	153
					Selection Probabilities				
IPW	$\mathbf{D_2}$	0.063	0.063	0.206	0.092	0.092	0.281	0.206	0.281
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.034	0.034	0.162	0.018	0.017	0.088	0.160	0.085
	$\mathbf{D_1}$ Net Study 2	0.053	0.119	0.178	0.104	0.228	0.168	0.234	0.281
ML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.130	0.394	0.394	0.001	0.745
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.206	0.524	0.448	0.088	0.800
CML	$\mathbf{D_2}$	0.001	0.001	0.001	0.361	0.361	0.836	0.001	0.836
	$\mathbf{D_1}$ Study 1	0.019	0.088	0.019	0.088	0.214	0.088	0.088	0.214
	$\mathbf{D_1}$ Study 2	0.001	0.001	0.001	0.125	0.398	0.398	0.001	0.760
	$\mathbf{D_1}$ Net Study 2	0.019	0.088	0.019	0.202	0.527	0.451	0.088	0.811

Appendix 4B: Additional Sequential Two-Phase Biomarker Studies in Psoriatic Arthritis

In this section, we report a study framed within the psoriatic arthritis (PsA) research project involving the University of Toronto Psoriatic Arthritis Cohort (UTPAC). We consider a dataset that involves a subset of the full registry of UTPAC comprising 1489 patients. One particular interest of the PsA research program is to assess the association between the biomarker matrix metalloproteinase 3 (MMP-3) and joint damage progression. However, it is not feasible to assay all the biospecimens to measure the MMP-3 levels of the entire cohort because of budgetary constraints. Similar to Section 4.5, patients have the conditions of their joints recorded, and the dataset provides auxiliary information such as gender and the erythrocyte sedimentation rate (ESR) levels for the entire cohort.

We design a focused simulation study to investigate the performance of our proposed secondary analyses and optimal sequential two-phase designs in the setting of the PsA program. According to the dataset, 251 patients have their MMP-3 levels measured at a baseline assessment to study the progression of clinical damaged joints in two years from the baseline. Such patients serve as pilot data to help with parameter configuration. In terms of sequential two-phase studies, we let $Y_1 = 1$ if a patient develops two or more clinically damaged joints of grade 1 (deformity) or higher in two years from the baseline, and $Y_1 = 0$ otherwise. We let the auxiliary covariate $Z = 1$ if the baseline ESR level is greater than 20 for females or greater than 13 for males, and $Z = 0$ otherwise. The exposure variable X is defined by dichotomizing the continuous MMP-3 levels according to its 0.8 quantile. As for the new response of interest in Study 2, it is defined by the increase in the number of tender and swollen joints in two years from the baseline. We let $Y_2 = 1$ if a patient develops two or more such joints during the two years. The response models and the covariates models have the form of (4.1), (4.2), (4.3), (4.4), and (4.10). Fitting logistic regression models to the data, we let $\alpha = (-1.727, 0.736, 0.554)'$, $\beta = (-2.258, 0.809, 0.144)'$, $\gamma = (-0.069, 1.322, 0.149, -2.565)'$, $\xi = (-1.893, 1.391)'$, and $\zeta = -0.055$.

We perform a simulation study with $nsim = 500$ and phase I sample size $n = 5000$.

Suppose that Study 1 employs BS based on 4 strata defined by Y_1 and Z to select a sub-sample of expected size $E(M_1) = 250$ for the measurements of the exposure variables. Without additional sampling, we perform joint secondary analyses to investigate $Y_2|X, Z$ using Y and Z of the entire dataset as well as X of the M_1 patients. Results are displayed in the top half of Table 4B1. Adopting joint response models, all the secondary analyses recover the parameter of interest well. Employing the marginal model of Y_2 via IPW as in (4.11) leads to valid results, too. We further consider a subsequent Study 2 with a phase II sub-sample of expected size $E(M_2) = 250$ to investigate the performance of our proposed optimal design $\mathbf{D_1}$ in the PsA setting. Results are displayed in the bottom half of Table 4B1. Our proposed optimal designs $\mathbf{D_1}$ give valid estimates with negligible EBias. The likelihood methods are found to be more efficient than the IPWEEs, illustrated by smaller ASEs and ESEs in the $\mathbf{D_1}$ rows. Hence, the proposed analyses and designs of sequential two-phase studies are well adapted to the PsA setting to inspect the relationship between the disease progression and biomarkers of interest with budget limitations.

Table 4B1: Empirical biases (EBias), average standard errors (ASE), empirical standard errors (ESE), and empirical coverage probabilities (ECP) of the parameter estimates of interest following the secondary analyses (top half) and the combined phase II data of the sequential two-phase designs (bottom half) using likelihood and inverse weighting methods in the PsA setting. For secondary analyses $E(M_1) = 250$, and for sequential two-phase studies $E(M_1) = E(M_2) = 250$. "BS" and "opt" stand for balanced sampling and optimal sampling, respectively. The response of Study 1 is whether a patients develops two or more clinically damaged joints of grade 1 or higher in two years of follow-up. The response of Study 2 is whether a patients develops two or more tender and swollen joints in two years of follow-up. $nsim = 500$, $n = 5000$, and $\beta_1 = 0.809$.

	Stratification			Results			
Response Model/Design	Study 1	Study 2	Analysis	EBias	ASE	ESE	ECP(%)
			Secondary analyses				
Joint	BS (Y_1, Z)		IPW	0.029	0.460	0.467	95.2
			IPW2	0.037	0.457	0.462	95.2
			ML	0.007	0.464	0.457	93.4
			CML	0.018	0.476	0.468	95.6
Marginal	BS (Y_1, Z)		IPW	0.019	0.463	0.468	95.8
			Sequential two-phase studies				
$\mathbf{D_1}$	BS (Y_1, Z)	opt (Y, Z)	IPW	0.021	0.257	0.262	93.8
			IPW2	0.021	0.257	0.262	93.6
			ML	0.019	0.245	0.244	96.4
			CML	0.019	0.247	0.249	95.6