IMPROVING INFERENTIAL AND COMPUTATIONAL EFFICIENCY FOR REAL-WORLD DATA

By

ELIZABETH ANNE SIGWORTH WESTERBERG

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

BIOSTATISTICS

December 17 th, 2022

Nashville, Tennessee

Approved:
Bryan Shepherd, Ph.D.
Qingxia “Cindy” Chen, Ph.D.
Ran Tao, Ph.D.
Jake Hughey, Ph.D.

To my husband, parents, sister, and friends.
Thank you for everything, I could not have done this without you.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Qingxia Chen, for providing me with invaluable guidance and support over the years. I have learned so much from working with you on this dissertation as well as through our collaborative projects. Thank you for making me a better researcher and a better statistician. Thank you also to my committee members, Dr. Bryan Shepherd, Dr. Ran Tao, and Dr. Jake Hughey for your feedback and insight, they truly helped shape the direction of this work.

I want to thank all of my collaborators in the Center for Precision Medicine, but most of all Dr. Jeremy Warner and Dr. Samuel Rubinstein. Jeremy, I am lucky to have been matched with you for my research assistantship so early in my time at Vanderbilt. Your passion for your research and dedication to your mentees never ceases to impress me, and I am very grateful to you for your guidance during the writing of my first research publication (and many thereafter!). And Sam, thank you for providing me with so many interesting research questions to tackle, and for your mentorship and guidance as I learned how to be an effective collaborator.

Next, I want to thank all of the professors who taught me how to think like a statistician while also being there to support me as a person. To Bob Johnson, for always letting me drop by your office to chat (sometimes even about statistics!), and for helping me finally get over my fear of probability theory. To Jeffrey Blume, for encouraging me to join the PhD program to begin with. You believed that I could do this even when I didn't and I can't thank you enough for that. And to my undergraduate mentors and professors at St. Olaf College, Kay Smith, Kathryn Ziegler-Graham, Sharon Lane-Getaz, and Kristina Garrett. I would not be where I am today without the foundation in mathematics and statistics that you made possible for me.

I was fortunate enough to secure an internship at a healthcare startup early in my graduate student career, and feel so lucky to have had those experiences to enrich my education. In particular, I want to thank all of the mentors that I've worked with over the years on various projects. Dana Blakemore, Amy Graves, Lindsey Clark, Ray Pasek, Mathilde Granke, Jay Arnett, and Maithili Rao, you all taught me how to translate my classroom education to real-world applications, and I hope you all know how much I appreciate that.

Next, I would like to thank my family, without whom I would not be where I am today. To my parents, Rob and Andrea Sigworth, thank you for all of the ways in which you have supported and encouraged me. Thank you for always answering the phone, for listening to me ramble about my research and always sounding interested even if you had no idea what I was talking about, and for buying us takeout from several states away when you knew we were too busy to cook. To my sister, Laura, you kept me supplied in memes and conversations about anything other than statistics, and I hope you know how much I appreciated those moments of normalcy whilst in the thick of work. Finally, my most heartfelt thanks and appreciation go to my husband, Jake. We started graduate school together early in our relationship, and have transitioned from dating to engaged to married in the time it took to finish this dissertation. It was no small feat to go through all of that together and end up where we are today. Thank you for always lending a listening ear, a proofreading eye, and an understanding heart. I cannot wait to take on our next adventure together.

I would be remiss not to thank all of the friends who have been there for me throughout the years. Kaitlyn, Madi, and Kelsie, not a lot of people can say that they've had the same best friends since the fourth grade, and I hope you know how dear you all are to me. Anna, our weekly workouts and coffee dates kept me sane

(and caffeinated), and I love that the distance from Minnesota to Tennessee couldn't get in the way of our friendship. I was also blessed to meet some of my dearest friends through this graduate program. Hannah, Josh (and Courtney, Ellie, Hazel, and Bowie), Nathan, Sarah, Valerie, and so many others, thank you for ensuring I never felt alone throughout this experience. And to all of my adoptive family in the Psychology department, Steven, Lisa, Brock, Nina, Jeff, Michelle, Micala, and everyone else I might have shared a lab beer with, you were all an essential part of my graduate school experience and I appreciate you all so much.

Finally, and deserving of the last paragraph, I want to thank our cats Boo and Clicquot. You have been along for the ride from the start, and have provided endless hours of distraction, companionship, and love. You were the perfect work from home office mates and made plenty of virtual friends over the last few years of Zoom calls. I love both of you from the bottom of my heart, even if you did bite my laptop screen and chew your way through more than a few charging cables. Please try not to chew up this dissertation, though, okay? I worked really hard on it.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

1.1 Abstract

Large clinical and epidemiological studies allow researchers to study drug and exposure effects across populations and increase the chance of observing rare events. However, with large studies come increased data volume, follow-up time, and overall complexity, necessitating analytical techniques that scale with increasing data sizes or perform efficient analyses on subsets or summaries. In this dissertation, we explored several facets of inferential and computational efficiency in the context of big data. First, we considered a case where complete data are unavailable due to data sharing limitations and developed a Bayesian meta-analysis model that uses clinical trial summary data to estimate equivalent dose pairs of two drugs based on adverse event rates. We demonstrated its efficiency when compared to models fit on individual subject data via simulation study, including under model misspecification. We then applied our method to clinical trial data on two taxane chemotherapy drugs known to cause peripheral sensory neuropathy. Next, we compared methods for estimating time-varying effects in the Cox model, evaluating via simulation five different options in terms of inferential and computational efficiency when applied to effects of varying complexity. We then applied the best performing methods to tumor survival data pulled from Vanderbilt University Medical Center electronic health records, using available ICD-9 codes and drug exposures to estimate the potential time-varying effect of metformin on survival in several tumor types. Finally, we applied the findings of our time-varying effect exploration to two-phase study designs, where inexpensive covariates are available for all participants, but an expensive covariate that is believed to have a time-varying effect is available only for a subset. This work extends on the previous development of a two-phase approach for the Cox proportional hazards model, and uses B-spline sieves to estimate the conditional density of expensive covariates given the available inexpensive covariates. We demonstrated its performance via simulation and comparison to existing methods that incorporate only Phase II data, and finished with an application to a large cohort study on the effects of oxidative stress over time on colorectal cancer development.

1.2 Introduction of Topics Covered

In this dissertation, we considered several aspects of inferential and computational efficiency when faced with large or complex data scenarios. As clinical and epidemiological studies are able to recruit and manage an increasing number of subjects (due to increased computing space, human resources, and international collaborative efforts), the resulting data generated by these studies correspondingly increases in size and potentially complexity. Before delving into our work, we briefly describe here the types of problems our methods looked to solve.

1.2.1 Building a Dose Toxo-Equivalence Model from a Bayesian Meta-Analysis of Published Clinical Trials

Clinical trials concern themselves primarily with the reporting of endpoints pertaining to the efficacy of the drug under study. Typically, these include summary statistics such as effect size estimates and odds ratios of survival or progression-free survival, among others (Whitehead, 2002). When synthesizing information across multiple trials in a meta-analysis, it is these estimates that are used to make inferences on the true value of the parameter of interest (Hedges and Olkin, 1985). However, researchers may be interested in other outcomes reported by trials in less detail, such as the incidence of adverse effects related to the drug under study. In our particular case, we were interested in two chemotherapy drugs, paclitaxel and docetaxel, prescribed to treat a wide variety of cancers. These drugs are both members of the taxane class of drugs known to induce peripheral sensory neuropathy as an adverse effect (Warner et al., 2015). This outcome, which can be debilitating and potentially permanent, is believed to be directly linked to cumulative exposure to taxanes (Argyriou et al., 2008). Additionally, clinicians often begin patients on paclitaxel as their first line of treatment, but some patients experience hypersensitivity or infusion reactions that necessitate switching to a docetaxel-based regimen. Despite the risk of neuropathy with exposure to taxanes, there is no clear guidance on choosing an initial dose and dosing frequency for docetaxel given the previous paclitaxel regimen, particularly in terms of a patient's overall risk of developing neuropathy.

As the rate of neuropathy is often reported by clinical trials as a secondary outcome, we sought to develop a method that could leverage this publicly available data to estimate dose-equivalence pairs of paclitaxel and docetaxel (or any two drugs of interest) that would be expected to produce equivalent neuropathy rates. While the ideal scenario would have been to collect individual patient data from each candidate trial, of which we identified 169, our efforts to contact corresponding authors and inquire about data access were unsuccessful. The majority of requests for contact received no reply, and those that did reply informed us that the data was not shareable due to either privacy concerns or restrictions put in place by the groups sponsoring the studies. Thus, we found ourselves in need of a method that relied only on publicly available data, but that could produce efficient results consistent with a model fit to individual subject data. We developed our model in a Bayesian framework and established its performance compared to individual subject data via simulation, then applied to the previously described taxane clinical trial meta data. The results of this work can be found in Chapter 2.

1.2.2 An Empirical Comparison of Methods for Efficiently Estimating Time-Varying Effects in the Cox Model

Next, we pivoted to a question involving time-to-event survival data. We were developing an analysis plan for the two-phase data in Chapter 4 (see below), which involved the estimation of a time-varying effect in the Cox model. However, while researching different methods for nonparametrically estimating a time-varying effect, we realized that there was no clear answer on what method under consideration would produce the most efficient results, both inferentially and computationally. As our eventual target dataset for this two-phase work was very large in scale ($> 64,000$ individuals), efficient estimation was important. So, we set out to compare five different methods for estimating a time-varying effect, $\beta(t)$, via both simulation studies and a real-world application. Four of our methods relied on basis functions for estimation, namely Bernstein polynomials, B-splines, penalized splines, and restricted cubic splines. The fifth, local linear estimation, used a kernel smoothing method to estimate $\beta(t)$ at pre-specified time points. We evaluated these methods using simulated data with increasingly complex time-varying effects, and fit each method to several tuning parameter specifications that controlled their flexibility and computational intensity. We then compared within methods and across methods based on their ability to estimate $\beta(t)$, and to do so efficiently while keeping computational needs as low as possible. Based on the results of these simulations, we applied several of the best candidate methods to a particularly large dataset pulled from Vanderbilt University Medical Center electronic health records (Roden et al., 2008), with $> 40,000$ subjects and

$> 2,400$ covariates. This analysis looked at the potential for a time-varying effect of metformin exposure on survival in individuals with several types of tumors, modeled while controlling for the demographic information, ICD-9 codes, and drug exposures that were available in the electronic health record. The results of these simulations and analysis are contained in Chapter 3.

1.2.3 Efficient Estimation of the Cox Model with Time-Varying Effects Under Two-Phase Designs

Finally, we tackled the problem of estimating time-varying effects in the Cox model under two-phase designs. We were motivated by data from a large cohort study, which looked at colorectal cancer incidence as it relates to oxidative stress. A previous pilot study had found an association between oxidative stress, as measured by a biomarker, and the eventual development of colorectal cancer, and this association was believed to be time-varying, switching from pro- to anti-carcinogenic over time. This seemed an ideal application for the results of Chapter 3, however additional work was needed to account for the type of data available from this study. More specifically, the cohort study performed two-phase sampling of its population, collecting all inexpensive covariates for all participants but only measuring levels of the biomarker for cases and a 1-1 nested case control matched group of controls. Thus, out of the original 64,410 subjects in this study, only 1,402 were included in the Phase II collection of biomarker information. Therefore, a method was needed that could account for this two-phase sampling and efficiently estimate the potential time-varying effect of biomarker levels.

We built on previous work by Tao et al. (2017), which developed an efficient two-phase estimation procedure for the Cox model under the proportional hazards assumption, using B-spline sieves to estimate the conditional density of the expensive covariate (in our case biomarker samples) given the available inexpensive covariates. In so doing, this method was able to incorporate all Phase I information as well as the Phase II information from those sampled. This is in contrast to other existing methods for analyzing two-phase data, such as inverse probability weighting or estimation methods specific to nested case control studies, which either exclude those from Phase I only (Støer and Samuelsen, 2012; Liu et al., 2010) or estimate the full likelihood but do so by making parametric assumptions about the association between the expensive and inexpensive covariates (Saarela et al., 2008). Our extension of the method of Tao et al. (2017) uses a B-spline to estimate the time-varying effect of the expensive covariate, denoted $\beta(t)$, and by incorporating all Phase I information was able to approach full efficiency when compared via simulation to models fit to complete information (where the expensive covariate was collected for all subjects). We compare the computational and inferential efficiency of our method to existing methods for time-varying coefficient estimation, and close with the performance of our method on the colorectal cancer data in Chapter 4.

CHAPTER 2

Building a Dose Toxo-Equivalence Model from a Bayesian Meta-Analysis of Published Clinical Trials

Elizabeth A.S. Westerberg, Samuel M. Rubinstein, Jeremy L. Warner, Yong Chen, and Qingxia Chen

2.1 Summary

In clinical practice, medications are often interchanged in treatment protocols when a patient negatively reacts to their first line of therapy. Although switching between medications is common, clinicians often lack structured guidance when choosing the initial dose and frequency of a new medication given the former with respect to risk of adverse events. In this paper we propose to establish this dose toxo-equivalence relationship using published clinical trial results with one or both drugs of interest via a Bayesian meta-analysis model that accounts for both within- and between-study variances. With the posterior parameter samples from this model, we compute median and 95% credible intervals for equivalent dose pairs of the two drugs that are predicted to produce equal rates of an adverse outcome, relying solely on study-level information. Via extensive simulations, we show that this approach approximates well the true dose toxo-equivalence relationship, considering different study designs, levels of between-study variance, and the inclusion/exclusion of non-confounder/non-modifier subject-level covariates in addition to study-level covariates. We compare the performance of this study-level meta-analysis estimate to the equivalent individual patient data meta-analysis model and find comparable bias and minimal efficiency loss in the study-level coefficients used in the dose toxo-equivalence relationship. Finally, we present the findings of our dose toxo-equivalence model applied to two chemotherapy drugs, based on data gathered from 169 published clinical trials.

2.2 Introduction

Often there are multiple medication regimens that can be prescribed to patients to treat the same type of illness. However, these regimens can differ in their dosing as well as in their risks of inducing adverse events in patients. As a motivating example, we concern ourselves with the taxane chemotherapy drugs paclitaxel and docetaxel, which are both known to induce peripheral sensory neuropathy, an outcome that is believed to be directly related to cumulative exposure. Patients are frequently switched from paclitaxel to docetaxel due to infusion reactions, yet there currently exists no clear guidance on how clinicians should choose an initial dosage and frequency of docetaxel given a patient's previous paclitaxel regimen. However, as with the side effects for many drugs in similar scenarios, the incidence rate of peripheral sensory neuropathy in clinical trials of paclitaxel and docetaxel is commonly reported in published literature. Thus, it is desired to develop a method that leverages this available pool of study meta-information to estimate the dose toxo-equivalence relationship.

Conventional meta-analysis approaches combine results from independent studies to find patterns or discrepancies in the published literature (Hedges and Olkin, 1985). They typically use summary statistics reported by each study, such as effect size estimates and standard errors, in either common effects or random effects models to make inferences on the true value of the parameter of interest (Hedges and Olkin, 1985). In the case of clinical trials, the reported summary statistics are often treatment effect estimates, such as odds ratios or risk differences between groups exposed to the two drugs of interest within the same study (Whitehead, 2002). As we are interested in the dose relationship and not the treatment effect, we need to extract the response rate for the treatment and its associated dosage information from each study rather than estimated treatment effect. Ultimately, we need a method to incorporate reported incidence rates for each drug at their specific dosage as well as potentially aggregated summary data from any study in one or both of these drugs, in a way that adds little bias, loses minimal efficiency, and produces a useful approximation of this dose toxo-equivalence.

The use of aggregated summary data in study-level meta-analysis has the potential to induce bias. Focused on treatment effect estimates, Berlin et al. (2002) investigated a real-world published example for which both individual patient-level and study-level data were available and ultimately recommended using individual patient data, when feasible, to study patient characteristics to avoid aggregation bias (in the presence of effect modifiers). In this paper, we will use the term individual patient data (IPD) meta-analysis for the regression analysis based on individual patients and refer to the analysis based on aggregated study-level data as study-level (SL) meta-analysis. When there is no interaction effect between patient characteristics and treatment, another question researchers asked was when IPD meta-analysis and SL meta-analysis would yield identical results. Among others, Olkin and Sampson (1998) and Steinberg et al. (1997) illustrated some special settings for which IPD meta-analysis and SL meta-analysis could generate identical treatment effect estimators, all assuming continuous outcomes in the IPD meta-analysis. When all covariates are at the study-level, SL and IPD analyses are generally equivalent. However, when there are individual-level factors, or when these factors are summarized at the individual level, IPD meta-analysis is preferred in practice over SL meta-analysis due to the risk of aggregation bias, particularly in the presence of an effect modifier (Berlin et al., 2002; Lambert et al., 2002). While IPD meta-analysis is preferred in these scenarios, oftentimes the IPD from different studies are difficult to obtain. Additionally, the size of the data used in IPD meta-analysis can lead to high computation time compared to SL meta-analysis, particularly with Bayesian posterior sampling. For these reasons, SL meta-analysis is more practical and feasible, especially when there exists no prior evidence of effect modifiers.

There is increasing interest in studying the relative efficiency of meta-analyses based on fitted results at the study level compared to IPD meta-analysis, given barriers on data sharing and/or protections on participant privacy. Among others, Lin and Zeng (2010b) showed that when compared to IPD meta-analysis approaches (called mega-analysis in their setting), common effect meta-analysis methods using effect estimates from models fit at the study-level have minimal efficiency loss, and Zeng and Lin (2015) further showed that random effect meta-analysis methods are at least as efficient as the former. However, their results do not apply to our setting. In particular, Lin and Zeng (2010b) considered a regression (see their first equation in Section 2.1) of an outcome Y_{ki} on covariates X_{ki}, for the i-th participant in the k-th study. Implicitly, they assumed covariates collected within each study (X_{ki}) had variations within k-th study, whereas in our setting, within a study, the dosage is fixed for the corresponding treatment (i.e., there is no variation within k-th study).

While various meta-analysis approaches have been intensively studied, little work has been done in the dose-equivalence model setting. The validity of SL meta-analysis under this setting has not been studied and its relative efficiency versus IPD meta-analysis has not been evaluated. We aim to fill this gap by developing a Bayesian random-effects model to study the dose toxo-equivalence relationship.

In this paper, we distinguish the covariates with no variation at the subject level (such as treatment and designed dosage) in the aggregated study-level data from those with variations (such as percentage of male or mean age) and call the former study-level covariates and the latter subject-level covariates. We propose a Bayesian random-effects SL meta-analysis model that accounts for both within- and between-study variances, with or without additional subject-level covariates, in Section 2.4. We show that under the proposed model the toxo-equivalence curve depends solely on the coefficients of the study-level covariates. Based on the posterior samples produced by this model, we compute median and 95% credible intervals for equivalent dose pairs of any two drugs of interest that are predicted to result in equal rates of the adverse outcome. Via extensive simulation studies in Section 2.5, we demonstrate the ability of this model to closely approximate the true dose toxo-equivalence relationship for different study designs, varying levels of between-study variance, and in the presence of subject-level data (which are not treatment or dose effect modifiers) in addition to study-level covariates. We compare the performance of this meta-analysis approach in terms of bias and efficiency to an IPD meta-analysis model fit on pooled subject-level information, and demonstrate that our method results in comparable levels of bias to the IPD approach, as well as minimal efficiency loss in all parameters used to calculate the dose toxo-equivalence relationship when the model is correctly specified. Additionally, we consider the sensitivity of our meta-analysis approach to various types of model misspecification. Finally, we illustrate our method with empirical data gathered from published clinical trials in either paclitaxel or docetaxel in Section 2.6. We conclude the paper with discussion in Section 2.7.

2.3 Motivating Example

Our motivating example looks at the chemotherapy medications paclitaxel and docetaxel, both members of the taxane class of drugs that are prescribed to treat a variety of cancers (Warner et al., 2015). Taxanes are known to induce peripheral sensory neuropathy, with patient risk for this outcome believed to be directly related to cumulative exposure (Argyriou et al., 2008). Clinicians frequently start patients on paclitaxel as their first line of therapy, but some are unable to continue treatment due to hypersensitivity or infusion reactions, at which point patients are often switched to a docetaxel treatment regimen. However, there is no clear guidance on how to choose the initial dose and schedule of docetaxel, given a previous paclitaxel regimen, particularly with respect to the overall risk of peripheral sensory neuropathy. Since the rate of neuropathy development within studies is commonly reported as an adverse effect of treatment, we performed a systematic review of randomized or non-randomized clinical trials of paclitaxel or docetaxel monotherapy among cancer patients aged $\geq$ 18 years, extracting all aggregated data necessary for the dose toxo-equivalence calculation. Individual patient data from the included studies was not attainable. We apply the method for study-level data described in Section 2.4 to this data in Section 2.6, after exploring its performance compared to individual patient data. Further insight into the clinical relevance of our approach, including complete details of the systematic review procedure and examples of our method applied to specific chemotherapy regimens can be found in our related clinical paper (Sigworth et al., 2022).

2.4 Methods

2.4.1 Hierarchical Model Structure

2.4.1.1 IPD Meta-analysis

We first consider the subject-level IPD meta model against which we will compare our SL meta approach. Denote $D_{ij} = (X_{iA}, X_{iB}, d_i, \mathbf{Z_{ij}})$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ as the subject-level covariates, where X_{iA} is an indicator that study i uses drug A, X_{iB} is an indicator that study i uses drug B, d_i is the dose received in study i normalized to mean 0 and standard deviation 1 (or a normalized version of a monotone transformation of the dose such as the square-root transformation), and $\mathbf{Z_{ij}}$ is a vector of subject-specific potential covariates assumed to be associated with the adverse event. Dose values were normalized to improve computational efficiency in the Bayesian fitting procedure (Kruschke, 2015). Additionally, N is the total number of studies and n_i is the number of subjects in study i. Let w_{ij} denote the incidence indicator of the adverse event of interest, with $w_{ij} = 1$ for a subject experiencing the adverse event and $w_{ij} = 0$ otherwise. We assume $w_{ij}|p_{ij} \sim Bernoulli(p_{ij})$ where without additional covariates we have

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i,$$

and with additional covariates we have

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i + \alpha_z' \mathbf{Z_{ij}}.$$

In this model, α_1 is the mean outcome for studies in drug A with a normalized dose of 0 and $\alpha_1 + \alpha_2$ is the mean outcome for studies in drug B with a normalized dose of 0. We also estimate a random intercept component for each study, μ_i, as a measure of between-study heterogeneity. Its variance, τ^2, measures between-study variance in responses not attributable to other included variables. Note that throughout this manuscript we use the subscript *i* to denote the study of interest, i.e. the group of subjects assigned to the same protocol within the same study. In this way, single-arm and multi-arm studies can be analyzed via this method. Noninformative priors of $\mu_i|\tau \sim N(0, \tau^2)$, with $\tau \sim InvGamma\ (0.001, 0.001)$ and $\alpha \sim MVN(\mathbf{0}, 10^6 diag(\mathbf{1}))$ are specified, where $\alpha = (\alpha_1, \ldots, \alpha_4, \alpha_z')$. Under this model, our posterior distribution for α, τ is

$$p(\alpha, \tau|\mathbf{D_{ij}}, w_{ij}) \propto p(w_{ij}|\mathbf{D_{ij}}, \mu_i, \alpha)p(\mu_i|\tau)p(\tau)p(\alpha).$$

2.4.1.2 Study-level Meta-analysis

Next, we propose a Bayesian hierarchical model of the prevalence of an adverse event across multiple studies using aggregated meta-data. Let Π_i represent the rate of the adverse event in study *i*, $\Pi_i = \frac{1}{n_i}\sum_{j=1}^{n_i} w_{ij}$, and define $Y_i = logit(\Pi_i)$. Denote the per-study data as $\mathbf{D_i} = (X_{iA}, X_{iB}, d_i, \mathbf{Z_i^a})$, where X_{iA}, X_{iB}, and d_i are as defined previously and are study-level variables taking the same value for all subjects in study *i*, and $\mathbf{Z_i^a}$ is a vector of aggregated summary statistics of $\mathbf{Z_{ij}}$ at each study *i*, such as frequencies of categorical variables or means of continuous variables. Then the outcome Y_i can be modeled as $Y_i|\mu_i, \beta, \mathbf{D_i} \sim N(\phi_i, S_i^2)$, where

$$\phi_i = \mu_i + \beta_1 + \beta_2 X_{iB} + \beta_3 X_{iA} d_i + \beta_4 X_{iB} d_i + \beta_Z' \mathbf{Z_i^a}.$$

Here S_i^2 is the within-study variance of our outcome, which depends on the size n_i of the relevant arm of the study as well as the count of adverse events in that outcome, k_i, such that $S_i^2 = 1/k_i + 1/(n_i - k_i)$ (the variance of a logit transformed proportion). As previously, β_1 represents the mean outcome for studies in drug A with a normalized dose of 0 and $\beta_1 + \beta_2$ is the mean outcome for studies in drug B with a normalized dose of 0, while μ_i represents a random intercept component estimated at the study level to measure between-study heterogeneity. We set noninformative priors of $\mu_i|\tau \sim N(0, \tau^2)$,

$\tau \sim InvGamma(0.001, 0.001)$, and $\beta \sim MVN(\mathbf{0}, 10^6 diag(\mathbf{1}))$, where $\beta = (\beta_1, \ldots, \beta_4, \beta_Z')$. Note that although noninformative priors were used in our simulations (and case study), informative priors could be considered if prior knowledge on the dose toxo-equivalence relationship was available. Based on this structure, the posterior distribution for β, τ is

$$p(\beta, \tau | \mathbf{D_i}, Y_i) \propto p(Y_i | \mathbf{D_i}, \mu_i, \beta) p(\mu_i | \tau) p(\tau) p(\beta).$$

When including additional covariates in the data-generating model, we consider SL meta-analysis models both with and without $\mathbf{Z_i^a}$, denoted SL-C and SL-NC respectively. The aggregated study covariates in the SL-C model are intended to adjust for the additional heterogeneity in responses that may be due to these values, but not to estimate α from the IPD meta-analysis. By considering both SL-C and SL-NC in the presence of additional covariates $\mathbf{Z_i^a}$, we can evaluate the sensitivity of our method to this extra information.

2.4.2 Equivalence Relationship

For our IPD meta-analysis fits, the hierarchical structure produces expected outcomes of

$$logit[E(w_{ij} | X_{iA} = 1, X_{iB} = 0, d_i = d_A, \mathbf{Z_{ij}} = \mathbf{z})] = \mu + \alpha_1 + \alpha_3 d_A + \alpha_z' \mathbf{z}$$
$$logit[E(w_{ij} | X_{iA} = 0, X_{iB} = 1, d_i = d_B, \mathbf{Z_{ij}} = \mathbf{z})] = \mu + \alpha_1 + \alpha_2 + \alpha_4 d_B + \alpha_z' \mathbf{z}$$

for studies in drugs A and B respectively, where the study in drug A had dose d_A and the study in drug B had dose d_B. From these expected outcomes, we can build a dose toxo-equivalence model for the dose and adverse outcome relationship between the two drugs of interest. As the logit is a monotone transformation, we look to solve

$$\begin{aligned} & logit[E(w_{ij} | X_{iA} = 1, X_{iB} = 0, d_i = d_A, \mathbf{Z_{ij}} = \mathbf{z})] \\ = \quad & logit[E(w_{ij} | X_{iA} = 0, X_{iB} = 1, d_i = d_B, \mathbf{Z_{ij}} = \mathbf{z})] \end{aligned}$$

with respect to d_B, assuming that d_A is known, which, when simplified, results in solving

$$d_{B,IPD} = (\alpha_3 d_A - \alpha_2)/(\alpha_4).$$

As for the SL meta-analysis, from our aggregated meta fits we can produce the expected outcomes for studies in drugs A and B as

$$\begin{aligned} E(Y_i|X_{iA} = 1, X_{iB} = 0, d_i = d_A, \mathbf{Z_i^a} = \mathbf{z}) &= \mu + \beta_1 + \beta_3 d_A + \beta_Z' \mathbf{z} \\ E(Y_i|X_{iA} = 0, X_{iB} = 1, d_i = d_B, \mathbf{Z_i^a} = \mathbf{z}) &= \mu + \beta_1 + \beta_2 + \beta_4 d_B + \beta_Z' \mathbf{z} \end{aligned} \tag{2.1}$$

where the doses in the studies are d_A and d_B respectively as before. We look to solve $E(Y_i|X_{iA} = 1, X_{iB} = 0, d_i = d_A, \mathbf{Z_i^a} = \mathbf{z}) = E(Y_i|X_{iA} = 0, X_{iB} = 1, d_i = d_B, \mathbf{Z_i^a} = \mathbf{z})$ with respect to d_B, which results in

$$d_{B,SL} = (\beta_3 d_A - \beta_2)/(\beta_4) \tag{2.2}$$

For each fit, we vary across a range of plausible values of d_A to find dose pairs (d_A, d_B) for which we expect the rates of adverse outcome to be equivalent, generating a posterior distribution for d_B. We use the Markov chain Monte Carlo methods employed by the JAGS software package in R to produce posterior samples for each of our model parameters. Let β_i^k be the kth MCMC sample from the posterior, then $d_{B,SL}^k = (\beta_3^k d_A - \beta_2^k)/(\beta_4^k)$.

We report the median and 95% credible intervals from our calculated distributions of $\{d_B^k, k = 1, \ldots, K\}$ with K being the total number of draws from the posterior. Similarly, we calculate dose pairs $(d_A, d_{B,IPD}^k)$ from the posterior of our IPD meta-analysis.

Note that under the additive model assumption in (2.1), the dose toxo-equivalence relationship in (2.2) is independent of the common intercept β_1 and the coefficients of the aggregated covariates, β_Z. This is critical because, as we will demonstrate in Section 4, there is minimal efficiency loss or increase in bias in estimating $(\beta_2, \beta_3, \beta_4)$ using SL meta-analysis when comparing to IPD meta-analysis, resulting in comparable dose toxo-equivalence relationship curves with the correctly specified model.

2.5 Simulation Study

2.5.1 Simulation for a Correctly Specified Model

2.5.1.1 Method

To evaluate the performance of our method in terms of bias, efficiency, and ability to recover the true dose toxo-equivalence relationship, we performed extensive simulations, varying study types (balanced and unbalanced), generating models (with or without subject-level covariates), and levels of between-study variability ($\tau^2 \in \{0, 0.5, 1\}$). We performed 500 repetitions for each combination, for a total of 6,000 simulations.

For each setting, we first simulated data from $N = 150$ study datasets, 75 in each drug. For consistency, we sampled 150 dose values from a $U(-1,1)$ distribution, which are used for each simulation repetition. Next, a random intercept $\mu_i, i \in \{1, \dots, 150\}$ is generated for each of the 150 studies from a $N(0, \tau^2)$ distribution. If the setting requires balanced studies, each of the 150 study-level datasets will contain $n_i = 100$ subjects; if unbalanced, each study contains between 50 and 200 subjects, in increments of 10, with equal probability. Each subject in study i is assigned the same X_{iA} and X_{iB} indicator variables based on the drug type under study in study i, i.e. if study i is in drug A then $X_{iA} = 1$ and $X_{iB} = 0$. If the study includes additional subject-level covariates, then each subject j has two subject-specific independent covariates drawn: a binary Z_{ij5} with study-specific probability θ_i drawn from $Unif(0.2, 0.5)$, and a continuous Z_{ij6} drawn from $N(0,1)$.

Once the subject-level covariate data is generated, we calculate subject-specific log-odds of the adverse event for subject j in study i with normalized dose d_i based on a pre-selected vector of study-level coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (-0.6, -0.8, -0.5, -0.9)$ and subject-level coefficients $(\alpha_5, \alpha_6) = (0.2, 0.5)$, the latter to be included in studies with additional covariates. These parameters were based on those estimated in our clinical application, discussed in Section 2.6, and produce non-rare event outcomes. Thus, for a generating model without additional covariates, the subject-specific log-odds are

$$logit(p_{ij}) = \mu_i - 0.6 - 0.8X_{iB} - 0.5X_{iA}d_i - 0.9X_{iB}d_i,$$

and for a generating model with two additional subject-level covariates, we have

$$logit(p_{ij}) = \mu_i - 0.6 - 0.8X_{iB} - 0.5X_{iA}d_i - 0.9X_{iB}d_i + 0.2Z_{ij5} + 0.5Z_{ij6}.$$

From $logit(p_{ij})$ we use the *expit* function to produce subject-specific probabilities of the adverse outcome, p_{ij}. We then generate adverse outcome indicators, w_{ij}, from a $Bernoulli(1, p_{ij})$. From this simulated data-set at the subject-level, we summarize to reflect the metrics commonly reported in a clinical trial. Dose d_i is the same across subjects within a study and does not need to be summarized. For binary covariate Z_{ij5} we take the mean across all subjects in study i, and for continuous covariate Z_{ij6} we take the median, creating study-level summary values Z^a_{i5} and Z^a_{i6}. Event rate per study is calculated as $\Pi_i = \frac{1}{n_i}\sum_{j=1}^{n_i} w_{ij}$. Studies with $\Pi_i \in \{0,1\}$ are dropped and a new dataset generated until 75 valid datasets in each drug have been created (fewer than 1 in 500 simulated studies were re-generated). Finally, the logit of the outcome, $logit(\Pi_i)$ is calculated, along with $\text{Var}(logit(\Pi_i)) = S^2_i = 1/k_i + 1/(n_i - k_i)$, where $k_i = \sum_{j=1}^{n_i} w_{ij}$.

Next, for the generating model without additional covariates, we fit our Bayesian models on both the aggregated study-level meta-data and the subject-level IPD meta-data. All models were fit using **JAGS** version 4.3.0 in **R** version 3.6.0 with packages **R2jags** version 0.6-1 (Su and Yajima, 2020) and **coda** version 0.19-2 (Plummer et al., 2006), and were executed using the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) cluster using 2 GB of RAM. Each model consists of four independent chains. Where $\tau^2 = 0$, we use a burn-in of 5000 and take 20000 samples with a thinning interval of 2, based on within-chain correlation from preliminary fits. For $\tau^2 \in (0.5, 1)$, we use a burn-in of 10000 and take 40000 samples with a thinning interval of 4. All R code necessary to reproduce these simulations can be found at https://github.com/esigworth/BayesianDTEM.

For the generating model with additional covariates, we fit three candidate models: an SL-C meta-analysis model including aggregated Z^a_{i5} and Z^a_{i6}, an SL-NC meta-analysis model fit on only study-level covariates, and the subject-level IPD meta-data model incorporating study-level covariates and subject-level Z_{ij5} and Z_{ij6}. Each of these models was fit using the same burn-in, sampling, and thinning settings as above based on τ^2.

From the full set of posterior samples of each parameter in each model we calculate the dose toxo-equivalence relationship across a range of normalized doses between -1 and 1, saving the median 95% credible interval bounds from each simulation. For model diagnostics, we report coverage probabilities, the ratio of the median absolute deviation (MADR) for each parameter in each SL meta fit to the MAD of the IPD meta fit, mean square error, percent bias, and relative efficiency. Percent bias is calculated as the median across 500 simulations of the difference between the mean of the posterior distribution for a given parameter and the true value of that parameter, divided by the true value and multiplied by 100. Relative efficiency is the median across simulations of the ratio of the variance in the posterior of the IPD model to the variance in the SL model (equivalent to the efficiency of the SL model over the efficiency of the IPD model, where efficiency is the inverse of variance). Both percent bias and relative efficiency are reported with 95% credible intervals.

2.5.1.2 Results

Our summary of the performance of this method focuses on the parameters involved in the calculation of the dose toxo-equivalence relationship, $(\beta_2, \beta_3, \beta_4)$ and $(\alpha_2, \alpha_3, \alpha_4)$, since it is only the performance of these parameters that will determine the ability of our meta and IPD meta models to approximate the true dose toxo-equivalence relationship.

We first assess our method in the absence of additional covariates. Table 2.1 summarizes the coverage probabilities and the MADRs and 95% credible interval widths (MADR CIW) for each parameter in each setting, arranged by study design and value of τ^2. Coverage is near 95% for all models and settings, and the MADR values for all fits are close to 1 with narrow MADR CIWs, signifying comparable levels of variability within the posterior sampling chains of our two model approaches.

In Figure 2.1 Panel **A** we display the median and 95% credible intervals for percent bias in parameter estimates from the SL meta and IPD meta model fits without additional covariates, with three levels of τ^2 $(0, 0.5, 1)$ displayed across the columns and study designs across the rows. Median percent bias is consistently close to 0, with variance around the median increasing with increasing τ^2. Figure 2.1 Panel **B** presents the median and 95% credible intervals for the relative efficiency of the SL meta model to the IPD meta model with no additional covariates. Efficiency is consistently close to 1, shifting slightly higher with

Table 2.1: Coverage, Median Absolute Deviation Ratios (MADR), Percent Bias (IQR), and MSE, by fit type and study design, for simulations without additional covariates.

Simulations Without Additional Covariates								
	Statistic	Coverage		MADR	Percent Bias (IQR)		MSE	
Setting	Fit (Parameter)	SL (β)	IPD (α)	SL (β) / IPD (α)	SL (β)	IPD (α)	SL (β)	IPD (α)
$\tau^2 = 0$								
Balanced	β_2 or α_2	0.96	0.96	1.00	-0.01 (0.67)	0.00 (0.69)	0.00	0.00
	β_3 or α_3	0.98	0.97	1.00	-0.02 (0.34)	-0.01 (0.34)	0.00	0.00
	β_4 or α_4	0.96	0.96	1.00	-0.01 (0.18)	0.01 (0.18)	0.00	0.00
Unbalanced	β_2 or α_2	0.96	0.96	1.00	-0.02 (0.61)	-0.01 (0.61)	0.00	0.00
	β_3 or α_3	0.94	0.94	1.00	-0.01 (0.35)	0.00 (0.35)	0.00	0.00
	β_4 or α_4	0.93	0.93	1.00	-0.01 (0.19)	0.00 (0.19)	0.00	0.00
$\tau^2 = 0.5$								
Balanced	β_2 or α_2	0.95	0.96	0.98	0.01 (2.27)	0.05 (2.35)	0.01	0.01
	β_3 or α_3	0.96	0.96	0.98	0.00 (1.22)	0.02 (1.23)	0.02	0.03
	β_4 or α_4	0.96	0.96	0.98	-0.02 (0.59)	0.00 (0.60)	0.02	0.02
Unbalanced	β_2 or α_2	0.96	0.96	0.98	-0.04 (2.15)	-0.01 (2.21)	0.01	0.01
	β_3 or α_3	0.97	0.97	0.98	-0.04 (1.11)	-0.02 (1.13)	0.02	0.02
	β_4 or α_4	0.93	0.93	0.99	-0.01 (0.64)	0.01 (0.66)	0.02	0.02
$\tau^2 = 1$								
Balanced	β_2 or α_2	0.96	0.96	0.98	-0.01 (2.81)	0.01 (2.89)	0.02	0.02
	β_3 or α_3	0.95	0.95	0.98	-0.06 (1.60)	-0.03 (1.66)	0.04	0.04
	β_4 or α_4	0.94	0.95	0.98	-0.04 (0.79)	-0.02 (0.81)	0.03	0.04
Unbalanced	β_2 or α_2	0.94	0.94	0.98	-0.06 (3.17)	-0.04 (3.24)	0.03	0.03
	β_3 or α_3	0.94	0.95	0.98	-0.05 (1.66)	-0.04 (1.71)	0.04	0.05
	β_4 or α_4	0.94	0.95	0.98	-0.03 (0.81)	-0.01 (0.83)	0.04	0.04

Table 2.2: Coverage, Median Absolute Deviation Ratios (MADR), Percent Bias (IQR), and MSE, by fit type and study design, for simulations with additional covariates.

Simulations With Additional Covariates												
	Statistic	Coverage			MADR		Percent Bias (IQR)			MSE		
Setting	Fit (Parameter)	SL-C (β)	SL-NC (β)	IPD (α)	SL-C (β) / IPD (α)	SL-NC (β) / IPD (α)	SL-C (β)	SL-NC (β)	IPD (α)	SL-C (β)	SL-NC (β)	IPD (α)
$\tau^2 = 0$												
Balanced	β_2 or α_2	0.96	0.95	0.97	0.98	0.98	-0.05 (0.65)	-0.06 (0.62)	0.01 (0.67)	0.00	0.00	0.00
	β_3 or α_3	0.89	0.88	0.95	0.98	0.97	-0.06 (0.35)	-0.06 (0.36)	0.00 (0.37)	0.00	0.00	0.00
	β_4 or α_4	0.78	0.76	0.94	0.98	0.98	-0.06 (0.20)	-0.06 (0.20)	0.00 (0.21)	0.00	0.00	0.00
Unbalanced	β_2 or α_2	0.96	0.95	0.96	0.98	0.98	-0.04 (0.58)	-0.05 (0.59)	0.00 (0.61)	0.00	0.00	0.00
	β_3 or α_3	0.91	0.90	0.98	0.98	0.97	-0.06 (0.29)	-0.06 (0.30)	0.00 (0.31)	0.00	0.00	0.00
	β_4 or α_4	0.73	0.72	0.95	0.98	0.98	-0.06 (0.16)	-0.06 (0.16)	0.00 (0.18)	0.00	0.00	0.00
$\tau^2 = 0.5$												
Balanced	β_2 or α_2	0.96	0.96	0.96	0.94	0.94	-0.12 (2.22)	-0.08 (2.23)	-0.03 (2.38)	0.01	0.01	0.01
	β_3 or α_3	0.96	0.96	0.95	0.94	0.93	-0.06 (1.19)	-0.07 (1.17)	-0.01 (1.25)	0.02	0.02	0.03
	β_4 or α_4	0.89	0.90	0.95	0.94	0.94	-0.08 (0.58)	-0.09 (0.59)	-0.02 (0.63)	0.03	0.03	0.02
Unbalanced	β_2 or α_2	0.95	0.95	0.95	0.94	0.94	-0.07 (2.23)	-0.08 (2.22)	0.00 (2.37)	0.01	0.01	0.01
	β_3 or α_3	0.96	0.95	0.97	0.94	0.94	-0.08 (1.11)	-0.08 (1.12)	-0.02 (1.18)	0.02	0.02	0.03
	β_4 or α_4	0.93	0.92	0.95	0.94	0.94	-0.07 (0.57)	-0.07 (0.58)	-0.01 (0.61)	0.02	0.02	0.02
$\tau^2 = 1$												
Balanced	β_2 or α_2	0.97	0.96	0.97	0.94	0.93	-0.03 (2.79)	-0.01 (2.83)	0.05 (3.05)	0.02	0.02	0.03
	β_3 or α_3	0.93	0.94	0.95	0.94	0.93	-0.06 (1.62)	-0.06 (1.63)	0.00 (1.72)	0.05	0.05	0.06
	β_4 or α_4	0.94	0.94	0.96	0.94	0.93	-0.09 (0.79)	-0.09 (0.80)	-0.02 (0.85)	0.04	0.04	0.04
Unbalanced	β_2 or α_2	0.95	0.94	0.95	0.94	0.94	-0.07 (2.99)	-0.07 (3.11)	0.02 (3.24)	0.02	0.03	0.03
	β_3 or α_3	0.93	0.93	0.94	0.94	0.94	-0.13 (1.63)	-0.13 (1.68)	-0.05 (1.77)	0.05	0.05	0.05
	β_4 or α_4	0.94	0.94	0.95	0.94	0.94	-0.08 (0.75)	-0.09 (0.74)	-0.02 (0.80)	0.03	0.03	0.03

increasing τ^2 but with credible intervals always containing 1, indicating comparable efficiency.

In Figure 2.1 Panel **C** we present the estimated dose toxo-equivalence curves and 95% credible intervals for the IPD meta (pink) and SL meta (blue) model fits with no additional covariates and compare these to the true dose toxo-equivalence relationship (green), by τ^2 (columns) and study design (rows). Credible interval widths increase with τ^2 in all settings. The SL and IPD estimated relationship curves are very similar to both the true model and one another across settings. With no additional covariates, then, the performance of the IPD meta and SL meta models in recovering the true dose toxo-equivalence relationship is comparable.

Next, we look at the performance of our method when additional subject-level covariates are included in the generating model, comparing the IPD meta fit on complete data to the SL-NC and SL-C models on study-level and potentially summarized data. Looking at Table 2.2, we have comparable coverage near 95% for β_2 in all model fits and for (β_3, β_4) where $\tau^2 \neq 0$. When $\tau^2 = 0$, we have low coverage of β_3 and β_4 in both the SL-C and SL-NC fits, between 70% and 90%. This may be due to $\tau^2 = 0$ being on the boundary of the $\mathit{InvGamma}(0.001, 0.001)$ prior on τ. The low coverage in this particular simulation setting is not a concern for our particular application, since it is highly unlikely for a collection of clinical studies to contain no between-study variability. All MAD ratios are just below 1, with narrow CIWs, showing that posterior sampling variability is still comparable.

Figure 2.2 Panel **A** shows the median and 95% credible intervals for percent bias in the parameter estimates from these three model fits. Percent bias is again at or very near 0, with variance in the percent bias increasing with increasing τ^2 for all considered parameters. In Figure 2.2 Panel **B** we see that both the SL-C and SL-NC fits are more efficient than the IPD fit for each parameter used in the equivalence calculation (as evidenced by the median relative efficiency being greater than 1 for each comparison), a trend that increases with τ^2.

Finally, Figure 2.2 Panel **C** shows the median and 95% credible intervals for the dose toxo-equivalence curves generated by each model fit, with IPD in pink, SL-C in blue, and SL-NC in orange. Each of these fall along the true curve (in green), and their credible intervals fully overlap on both sides of the estimated median, with their width increasing with τ^2. The inclusion of summarized subject-level information $\mathbf{Z_i^a}$ in the SL-C model does not improve the estimation of the dose

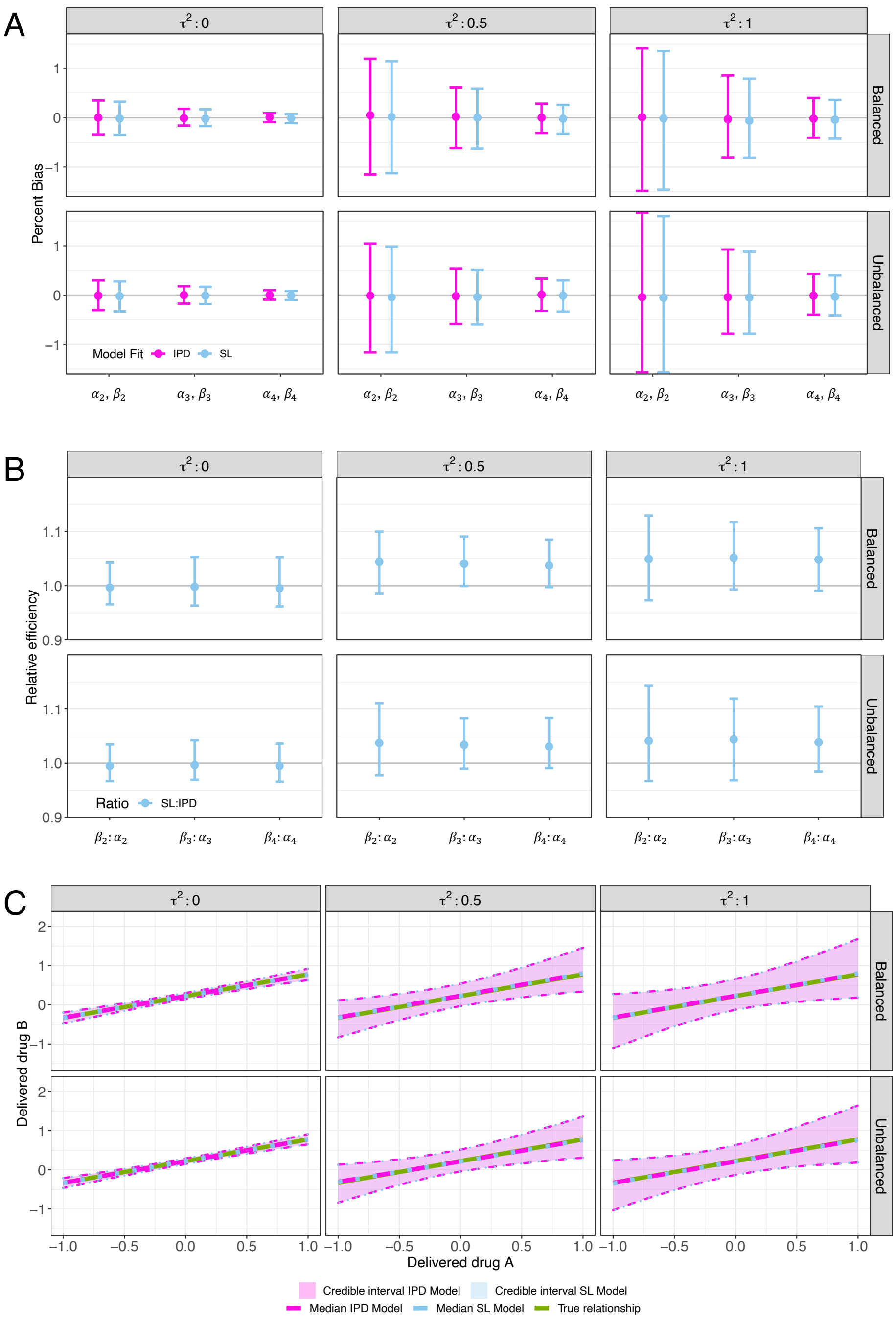

Figure 2.1: Summaries of correctly specified simulations without additional covariates, under each study design and value of τ^2. **A**: Percent bias and 95% credible intervals. **B**: Efficiency and 95% credible intervals of SL versus IPD meta models. **C**: Estimated dose toxo-equivalence relationships and 95% credible intervals of meta and IPD meta models compared to known relationship

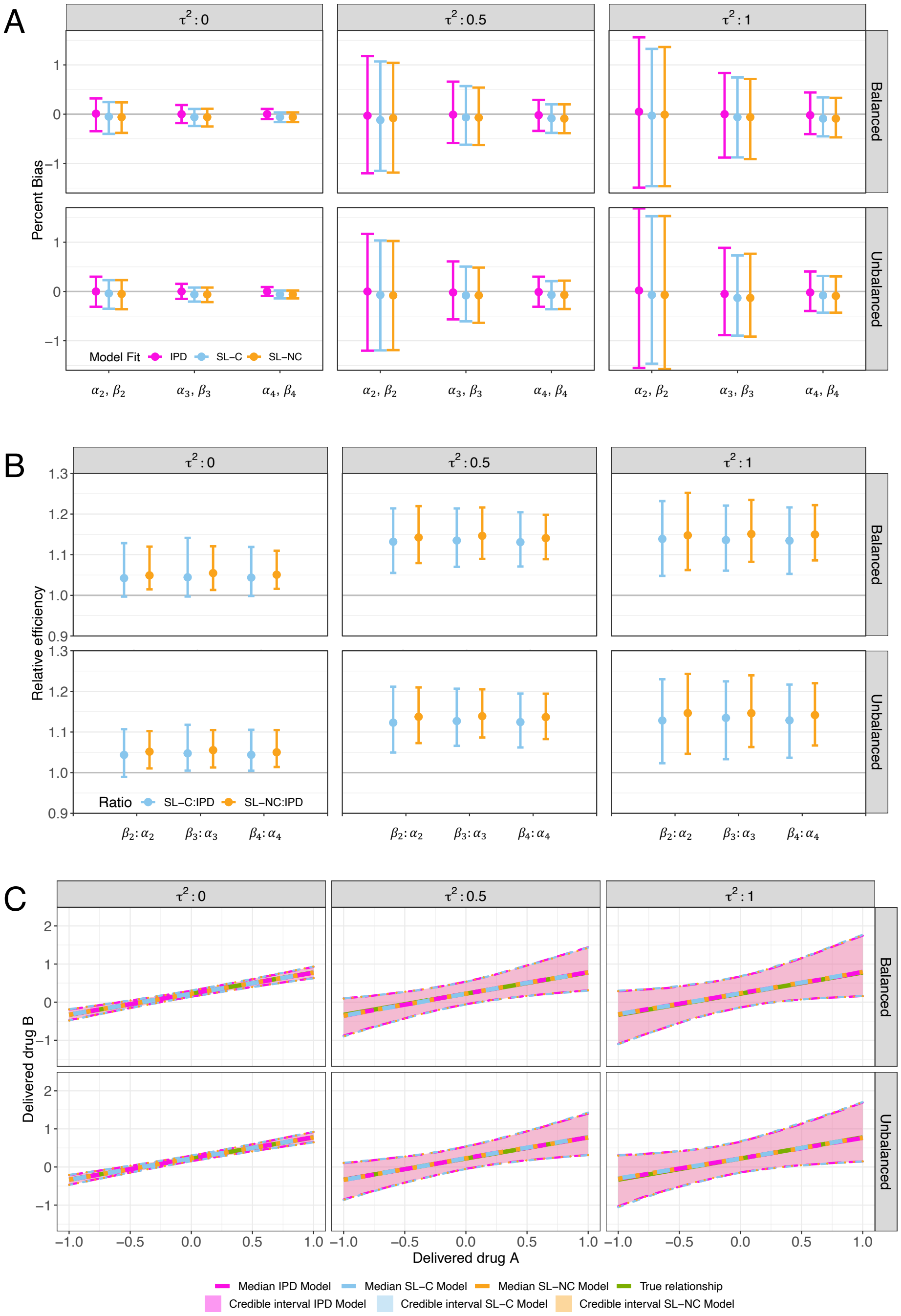

Figure 2.2: Summaries of correctly specified simulations with additional covariates, under each study design and value of τ^2. **A**: Percent bias and 95% credible intervals. **B**: Efficiency and 95% credible intervals of SL-C and SL-NC models versus IPD meta models. **C**: Estimated dose toxo-equivalence relationships and 95% credible intervals of meta and IPD meta models compared to known relationship

toxo-equivalence curve compared to the SL-NC model.

Of note is the reduced computing time for the SL model comparing to the IPD model. For example, with a single core of an Intel Sandy Bridge architecture processor using 2 GB of system memory, it took one minute to fit an SL model and five hours to fit an IPD model for a balanced study with no additional covariates and $\tau^2 = 1$. With additional covariates, the SL-C and SL-NC models again took one minute, while the IPD model took about 24 hours. Additionally, to check the sensitivity of the method to the simulation coefficients, an alternative set of coefficients chosen to produce similar event rates, $(\alpha_1, \ldots, \alpha_6) = (-0.4, 0.1, 0.6, 0.2, 0.3, 0.8)$, were also tested to assess the sensitivity of the model to coefficient choice, and produced results consistent with the above simulations in terms of agreement between the IPD and SL/SL-C/SL-NC models.

2.5.2 Simulation Under a Misspecified Model

2.5.2.1 Misspecified Model Simulation

We conducted the simulations in Section 2.5.1 assuming a correctly specified model. There are several misspecification scenarios that are of interest in evaluating the performance of our proposed method in practice. For example, it is possible that individual patient characteristics such as age, sex, and related comorbidities (Lutz et al., 2001) influence their response to medications or general risk for adverse events, possibly in a nonlinear fashion. When this is the case, it is necessary to have complete subject-level data to be able to accurately model these interactions (Debray et al., 2015). However, of interest is how closely we can approximate the dose toxo-equivalence relationship in the absence of subject-specific information. We explore this first via the inclusion of a nonlinear effect of Z_{ij6}, and then via an interaction between Z_{ij5} and either drug type or dose in the data-generating model, comparing the correct specification of the IPD model to the SL-C and SL-NC models fit as specified in Section 2.5.1 in the context of aggregated covariates (excluding interactions).

Additionally, our previous simulations assume that all subjects in study *i* receive the same dose. In practice, however, dosing among subjects can vary due to adverse reactions, missed appointments, or subject non-compliance (Lebovits et al., 1990). When this is the case, studies will often report an aggregated dose variable such as the median and IQR of received doses, as in our application in Section

2.6. To assess the sensitivity of our method to the accuracy of reported doses, we allow for variability in the received dose at the IPD level, and fit the SL model on an aggregated dose measure, in this case the median normalized dose. For complete simulation details for these four misspecification scenarios, see the Appendix.

2.5.2.2 Misspecified Model Results

Figure 2.3 shows the estimated dose toxo-equivalence curves for the simulations where subject-level covariates influence the risk of an adverse event in ways that are not considered at the SL-C and SL-NC levels. First, in Panel **A**, we allow the effect of the continuous covariate Z_{ij6} to be nonlinear. While the nonlinear term is only accounted for in the IPD model, we find that both the SL-C and SL-NC models are still able to approximate the true toxo-equivalence relationship as evidenced by all of the curves overlapping with the true relationship (in green).

Next, in Panels B and C we consider our simulations where Z_{ij5} is a mediator of the drug or dose response. We treat the binary covariate Z_{ij5} as sex, with $Z_{ij5} = 0$ indicating male and $Z_{ij5} = 1$ indicating female. We display the estimated IPD curves for females (green) and males (yellow), the estimated SL-C (pale blue) and SL-NC curves (orange), and the true curves for females (red) and males (dark blue). In both cases, we find the IPD meta model is fairly good at estimating the sex-specific equivalence curves, while the SL-C and SL-NC curves are roughly equal to one another and between the sex-specific curves. Thus, the SL-C and SL-NC curves provide a reasonable estimate of the relationship across sexes, but not at a sex-specific level.

Finally, the performance of our model when the dose truly varies at the subject level can be found in Figure 2.4. When allowing dose to vary by subject, we see an increase in variance of the percent bias of β_3 across all simulations (Panel **A**) as compared to the correctly specified models with no covariates in Figure 2.1 Panel **A**, however median percent bias is still at or very near to 0. Additionally, the correctly specified IPD meta model is more efficient than the misspecified meta model for β_3 and β_4 (Panel **B**). Finally, the credible interval for the meta model in the dose toxo-equivalence relationship (Panel **C**) is slightly wider than the IPD meta model, though the median estimated curves are still very similar to both one another and the true relationship.

2.6 Real-World Application: Taxane Treatment and Neuropathy

We revisit our clinical application from Section 2.3 by applying our methods to data from 169 published studies in the taxane chemotherapy drugs paclitaxel and docetaxel. Descriptive summaries of our considered studies can be found in our companion manusript (Sigworth et al., 2022). Our outcome Π_i is the observed rate of all-grade neuropathy in each study, and our dose d_i in each study is the normalized median cumulative dose received by subjects, calculated as $d_i = (D_i - \overline{D_i})/(sd_i)$ where D_i is the median cumulative dose received in mg/m^2 in study i, $\overline{D_i}$ is the mean of the reported D_i, and sd_i is the standard deviation of the reported D_i. Dose was normalized within each drug independently (i.e. paclitaxel was normalized with respect to other paclitaxel doses, and the same for docetaxel). We consider the inclusion of the study summary variables normalized median age, age_i, and dose gap, dg_i (time between taxane doses in fractions of four-week periods, i.e. 0.25 = 1 week).

We considered several transformations of D_i prior to normalization, looking to reduce skewness in the distribution of doses. Specifically we looked at $\sqrt{D_i}$, $\ln(D_i)$, and Box-Cox transformations with $\lambda = 0.22$ (chosen to maximize the objective function), $\lambda = 0.25$ (for fourth roots), and $\lambda = 0.33$ (for cube roots, as D_i is a volume). We fit both SL-C and SL-NC models, and considered the addition of a random slope to allow the dose-response relationship to vary by study. Each candidate model was fit with a burn-in of 15,000 and 500,000 samples with a thinning interval of 50, across four independent chains, and the deviance information criterion (DIC) was used to compare across models. The DIC values for all models (7 in total) were between 206 and 208 (listed in full in Supplementary Table 2), but the lowest and most parsimonious was the fit with covariates, no random slope, and no transformation to dose prior to normalization. Thus, the overall final model structure fit to the data can be defined as follows. Let $\mathbf{X_i} = (X_{iP}, X_{iD}, d_i, age_i, dg_i)$ be a study-level data vector, and define $Y_i|\mu_i, \beta, \mathbf{X_i} \sim N(\psi_i, S_i^2)$, where β denotes the vector of estimated coefficients based on our real data and

$$\psi_i = \mu_i + \beta_1 + \beta_2 X_{iD} + \beta_3 X_{iP} d_i + \beta_4 X_{iD} d_i + \beta_5 age_i + \beta_6 dg_i$$

with $X_{iP} = I(drug = paclitaxel)$ and $X_{iD} = I(drug = docetaxel)$. Noninformative priors of $\mu_i|\tau \sim N(0, \tau^2)$, $\tau \sim InvGamma(0.001, 0.001)$, and $\beta \sim MVN(\mathbf{0}, 10^6 diag(\mathbf{1}))$ were set for all model parameters. With the added normalization step for dose D_i, the equivalence relationship for dose D_P of paclitaxel to D_D of docetaxel, in mg/m^2, becomes

$$D_D = \frac{\beta_3\left(\frac{D_P-\overline{D_P}}{sd_P}\right) - \beta_2 + \beta_4\left(\frac{\overline{D_D}}{sd_D}\right)}{\beta_4/sd_D}$$

where $\overline{D_P}$ is the mean cumulative dose of paclitaxel, $\overline{D_D}$ is the mean cumulative dose of docetaxel, and sd_P and sd_D are the scaling values for paclitaxel and docetaxel respectively. Once the equivalence relationship was established, the resultant equivalent dose pairs were converted from normalized units to original mg/m^2 units using the centering and scaling values originally used in their normalization.

The final model diagnostics can be found in Figure 2.5 Panels A and B. Each parameter had convergence in the sampling chains, demonstrated by the Gelman plot in Panel **A**. There was no evidence of autocorrelation issues in the samples for $(\beta_2, \beta_3, \beta_4, \beta_5)$, demonstrated in the ACF plot for the β parameters in Panel **B**. There is some evidence of autocorrelation up to a lag of 40 with (β_1, β_6); however, as the chain is thinned by 50 and these parameters are not used in the equivalence calculation, this autocorrelation is not an issue.

The dose toxo-equivalence relationship generated from these samples can be found in Figure 2.5 Panel **C**, evaluated along a range of plausible cumulative paclitaxel doses in mg/m^2. Compared to the simulated curves, the width and shape of the credible interval around this relationship is consistent with our simulation with additional covariates, an unbalanced study design, and higher between-study variability, which is also the most similar simulation design given the mean estimate for $1/\tau^2$ (an estimate of precision returned by JAGS) was 1.12, resulting in a τ^2 of 0.89. Note the lower credible interval is clipped at 0 mg/m^2, since dosage values must be non-negative. Along the IQR of paclitaxel doses observed in our data, $656 - 1085 mg/m^2$, the width of the credible interval was roughly equal to the cumulative dose of two treatment cycles of docetaxel, providing useful guidance to clinicians. The lower bound of the credible interval is also particularly informative in a clinical sense, since clinicians can view the lower bound as a cautious dosing threshold from a toxicity perspective. Although we had also hoped to control for cancer type due to different proportions of each cancer represented for each drug, due to sample size constraints we were unable to do so. As more trial data becomes available, we hope to repeat this analysis while controlling for cancer type.

2.7 Discussion

This proposed approach to building a dose toxo-equivalence model from a Bayesian SL meta-analysis consistently produced a good approximation of the true dose toxo-equivalence relationship, performing very similarly to the IPD meta model fit to the full data under all simulation conditions with significantly reduced computing time (e.g. about one minute for the SL meta model as opposed to hours for the IPD meta model). Of note is that, in the absence of an interaction, the relationship between d_A and d_B depends solely on study-level information. This finding is valuable, since efforts to recover individual patient data with clinical trial results are time-intensive, expensive, or even impossible. In estimating the parameters involved in the dose toxo-equivalence relationship, we see no increase in bias or loss of efficiency across conditions as compared to IPD meta-analysis for those parameters needed to calculate equivalence, an outcome that has been demonstrated in similar works (Luo et al., 2022; Zeng and Lin, 2015). In the case of Zeng and Lin (2015), our approach has two major differences to their work. First, Lin and Zeng (2010b,a); Zeng and Lin (2015) were interested in using meta-analysis to combine the parameter estimates from each study/site, with each study including both groups and the parameter being the comparison of two groups, such as an odds ratio, while we are interested in incidence rates in each treatment group with specific doses, and each study can include either one or both groups. Secondly, adjusting for covariates has a different meaning in their application. In Lin and Zeng (2010b,a); Zeng and Lin (2015), the subject-level covariates were adjusted for at the study level before proceeding to meta-analysis, while in our application, only the aggregated study-level covariates are available, but not the subject-level covariates. In light of these differences, the theoretical conclusions made by Lin and Zeng are not applicable to our empirical findings.

We found no significant improvements in the dose toxo-equivalence estimates when including additional available aggregated covariates in our SL-C model fits, suggesting that potential heterogeneity that may be explained by these aggregated measures does not provide additional information in the estimation of equivalence, likely because these parameters are not estimating the same quantity between the SL meta and IPD meta models; in extreme cases these parameters could suffer from Simpson's Paradox (Cates, 2002; Berlin et al., 2002). Furthermore, the finding of comparable performance between SL-C and

SL-NC was in the absence of an effect modifier. As orthogonality between predictors does not necessarily lead to orthogonality of their coefficient estimates under a logistic regression model (McCullagh and Nelder, 1989), this finding is an empirical observation based on our simulation results. In practice, we could consider both SL-C and SL-NC approaches followed by model selection with, however, no intention to draw connections between the effects of aggregated covariates $\mathbf{Z_i^a}$ and the effects of subject-level covariates $\mathbf{Z_{ij}}$.

Our explorations in Section 2.5.2 demonstrate that our method is robust to several common types of model misspecification. When a continuous covariate has a nonlinear effect that is overlooked in the SL approach, our method still performs comparably to the IPD model with the appropriate specification. In the case where the drug or dose response is moderated by a binary covariate at the subject level, our method produced an equivalence curve that fell between the curves produced by the two levels of the binary covariate, providing a reasonable approximation of the average relationship in the face of incomplete information. Given that drug and dose responses frequently differ by sex or the presence of comorbid conditions (Lutz et al., 2001), this suggests that our method can still provide a useful guide for the dose toxo-equivalence in this context, though there is some loss of information in this context when using an SL-C or SL-NC model as opposed to an IPD model. Further work would be needed to assess the performance of our method in the case of an interaction with a continuous covariate, such as age or a lab value. Additionally, our simulations did not consider the inclusion of an interaction term in the SL model (with the aggregated $\mathbf{Z_i^a}$); the performance of this model specification is of interest in future investigations. We also found that our model was robust to some variation in the dose received at the subject level, explored at a variation of 10%. This finding supports our model application to paclitaxel and docetaxel, which used the median cumulative dose across each study. However, additional simulations with increasing dose variability could provide more insight into the impact of using an aggregated dose value. Finally, all of the models considered in the simulations and case study assume a linear relationship between transformed dose and rates of neuropathy. An extension on our work would be to explore model performance where the true relationship is non-linear, which would require constraints on the dose-equivalence calculation in order for the solution to be identifiable.

Of note is that our simulations and case study focus on non-rare outcomes. The taxane clinical trial data we used had an overall median adverse outcome rate of 0.29 (IQR 0.16 - 0.48), and the parameters used in our simulations generated similar outcome rates for comparability. Some refinement of our proposed method may be needed to be applicable to studies with rare outcomes, such as the use of integrated likelihood inference (Severini, 1998; Berger et al., 1999), a generalized linear mixed model as opposed to a linear mixed model at the study level to avoid the normal approximation, or the incorporation of Poisson random effects models (Cai et al., 2010). Additionally, a different prior for the random intercept term μ_i may be helpful in the case of rare outcomes, such as uniform or half-t, as the inverse-gamma prior may incorrectly weight variances when data is sparse (Gelman, 2006).

Acknowledgments

Thanks to Leena Choi for discussion on the different misspecification scenarios that we explored. This work was conducted in part using the resources of the ACCRE cluster at Vanderbilt University, Nashville, TN.

Funding

- E.A.S.W. was supported by NIH grant U24CA194215.
- S.M.R. was supported by NIH grant T32HG008341.
- J.L.W. was supported by NIH grant U24CA194215.
- Y.C. was supported in part by NIH grants R01LM012607 and R01AI130460.
- Q.C. was supported by NIH grants U24CA194215, R01CA237895, and R01CA188214.

Appendix: Misspecification Details

As described briefly in Section 2.5.2, we conducted simulations under four misspecification scenarios to test the sensitivity of our method. For the first, we allow for the subject-specific continuous covariate Z_{ij6} to have a nonlinear effect. In the next two, we consider the possibility that the subject-specific covariate $\mathbf{Z_{ij}}$ is a moderator of the dose toxo-equivalence relationship. Finally, we assess the sensitivity of our method to the accuracy of the reported dose value d_i, since there may be unreported variation in total dose received across study subjects. In this supplement, we provide more details on the specification of each of the models considered.

To start, we allow the continuous covariate Z_{ij6} to have a nonlinear effect in the data-generating model, where

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i + \alpha_5 Z_{ij5} + \alpha_6 Z_{ij6} + \alpha_7 (Z_{ij6})^2.$$

The previously specified values of $(\alpha_1, \ldots, \alpha_6)$ are used as in Section 2.5.1, and $\alpha_7 = 0.3$ The IPD model is fit with the correct nonlinear specification, while the SL-C and SL-NC models are fit following Section 2.5.1 in the context of aggregated covariates (with no inclusion of a nonlinear effect). For this and the following misspecifications, we fit 100 repetitions, considering balanced and unbalanced designs as well as $\tau^2 \in (0, 0.5, 1)$, for a total of 600 simulations, using the same sampling specifications as in Section 2.5.1.

Next, we consider the cases where the covariate $\mathbf{Z_{ij}}$ is a moderator of the drug response by adding an interaction term with either drug type or dose in the data-generating model. To do this, we consider two new data-generating models. In the first, we add an interaction between drug type and our binary subject-specific covariate, Z_{ij5}, such that

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i + \alpha_5 Z_{ij5} + \alpha_6 Z_{ij6} + \alpha_7 X_{iB} Z_{ij5}.$$

We use the same values of $(\alpha_1, \ldots, \alpha_6)$ as in Section 2.5.1, and set $\alpha_7 = 0.3$. Our IPD model is fit with the correct interaction specification. The SL-C and SL-NC models are fit as in Section 2.5.1 in the context of aggregated covariates (excluding interactions). The equivalence relationships under our SL-C and SL-NC fits remain unchanged from Section 2.4.2, while the correctly specified IPD fit produces separate equivalence curves for $Z_{ij5} = 0$ and $Z_{ij5} = 1$.

Similarly, we consider an interaction between dose d_i, treatment type (X_{iA}, X_{iB}), and Z_{ij5}, where

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i + \alpha_5 Z_{ij5} + \alpha_6 Z_{ij6} + \alpha_7 X_{iA} Z_{ij5} d_i + \alpha_8 X_{iB} Z_{ij5} d_i.$$

We again use the $(\alpha_1, \ldots, \alpha_6)$ from Section 2.5.1, and set $\alpha_7 = -0.3$ and $\alpha_8 = 0.4$. The IPD model has the correct specification, and the SL-C and SL-NC models are fit as in Section 2.5.1. The equivalence relationships for SL-C and SL-NC are unchanged from Section 2.4.2, and the IPD model produces an equivalence curve for each value of Z_{ij5}.

The final misspecification of interest is the sensitivity of our method to the accuracy of the dose value d_i. Our previous simulations assume that all subjects in study i receive the same dose. In practice, however, dosing among subjects can vary due to adverse reactions, missed appointments, or subject non-compliance (Lebovits et al., 1990). When this is the case, studies will often report an aggregated dose variable such as the median and IQR of received doses, as in our application in Section 2.6. To simulate this misspecification, we treat the original sampled doses from a $Unif(-1,1)$ as the "target" dose for each study, d_i, and then define a "high dose" as 10% greater than this target, $1.10d_i$, and a "low dose" as 10% less, $0.9d_i$. Each subject in study i receives one of these three doses with equal probability, denoted d_{ij}, and their outcome is generated as

$$logit(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_{ij} + \alpha_4 X_{iB} d_{ij}.$$

For the SL model, the dose for each study is the median of all doses d_{ij}, $j = 1, \ldots, n_i$, denoted $\widetilde{d_i}$. The model is then fit as specified for the SL meta model without covariates in Section 2.5.1, replacing d_i with $\widetilde{d_i}$. The covariates $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ from Section 2.5.1 are used.

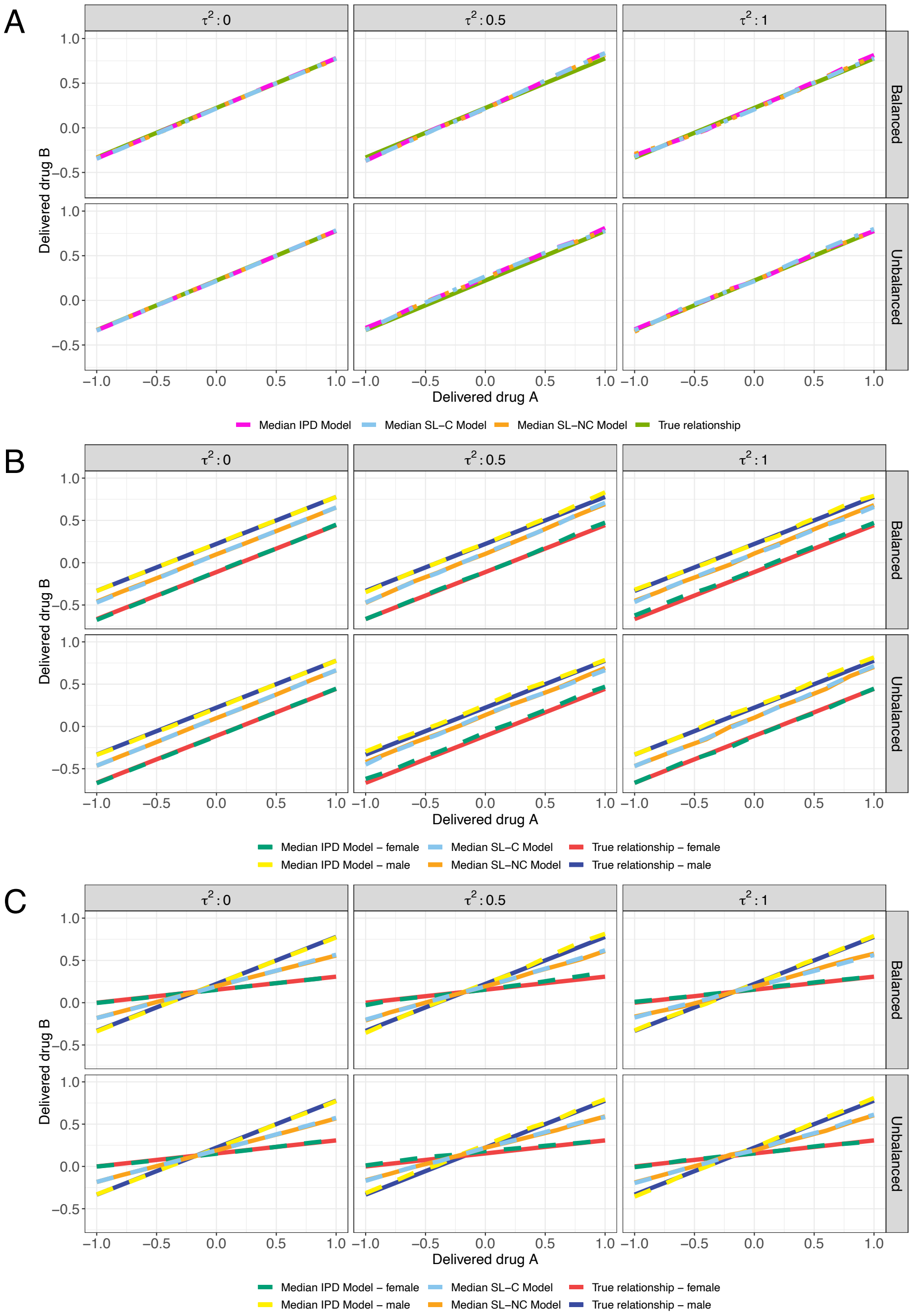

Figure 2.3: Estimated dose toxo-equivalence relationships and 95% credible intervals of SL meta and IPD meta models compared to known relationship with considered excluded term misspecificiations for each study design and value of τ^2. **A** Exclusion of nonlinear effect of continuous subject-level covariate. **B** Exclusion of interaction between drug type and a binary covariate (sex). **C** Exclusion of interaction between dose and a binary covariate (sex).

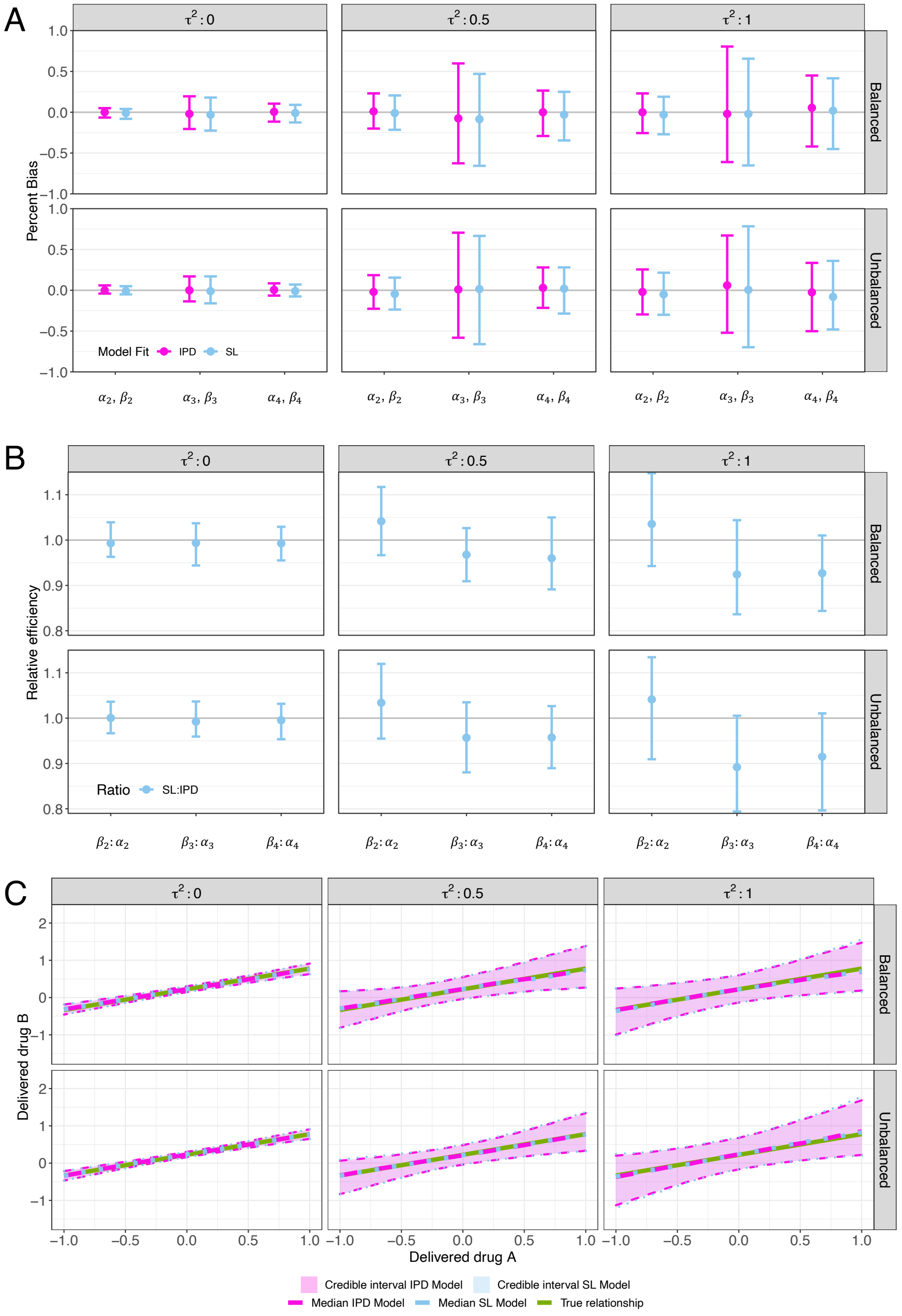

Figure 2.4: Summaries of misspecified simulations with a varied dose, under each study design and value of τ^2. **A**: Percent bias and 95% credible intervals. **B**: Efficiency and 95% credible intervals of SL versus IPD meta models. **C**: Estimated dose toxo-equivalence relationships and 95% credible intervals of meta and IPD meta models compared to known relationship.

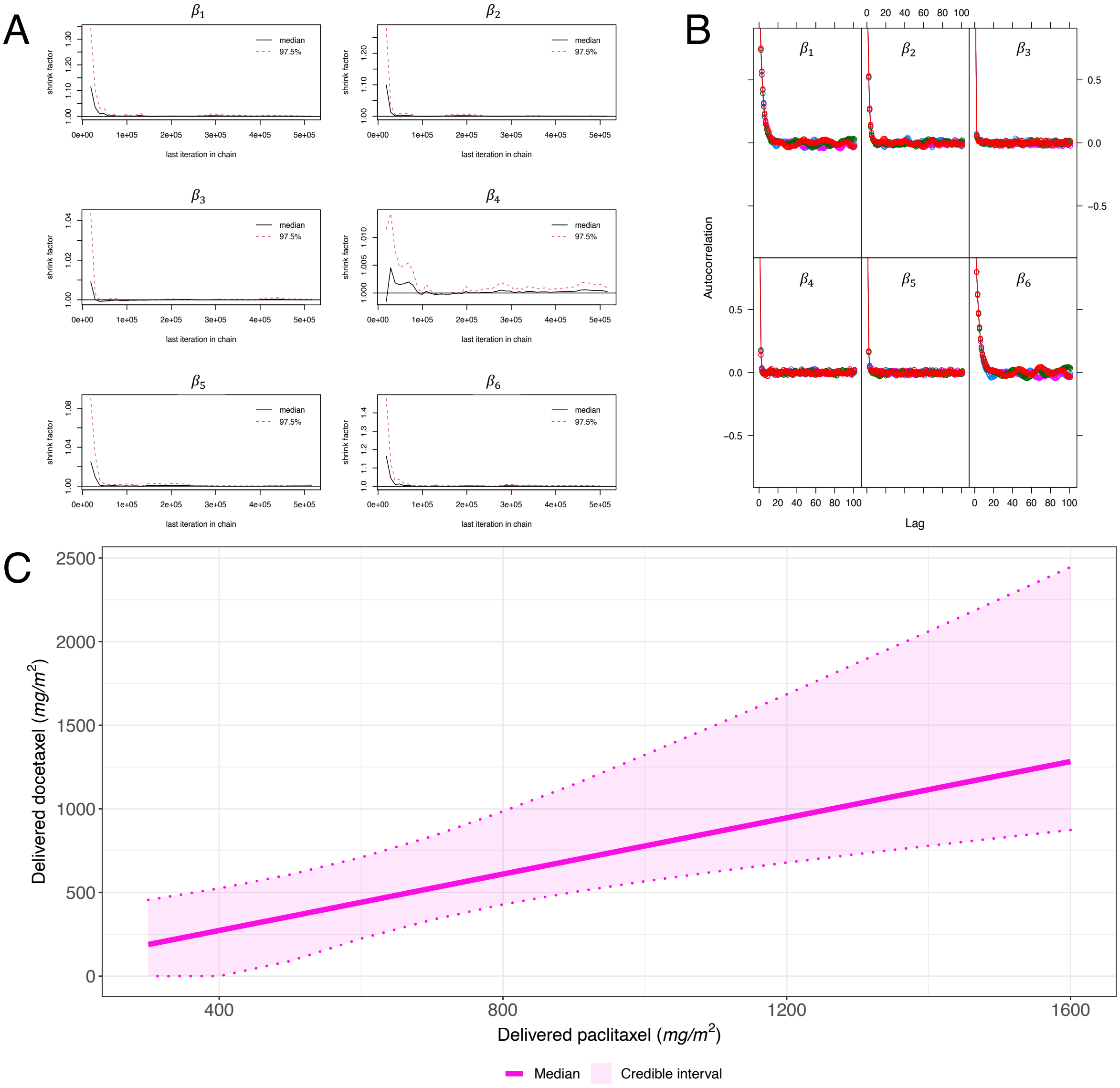

Figure 2.5: Summaries of model fit to taxane data. **A**: Gelman plot. **B**: ACF plot. **C**: Estimated dose toxo-equivalence relationships and 95% credible intervals for paclitaxel and docetaxel.

CHAPTER 3

An Empirical Comparison of Methods for Efficiently Estimating Time-Varying Effects in the Cox Model

Elizabeth A.S. Westerberg, Ran Tao, and Qingxia Chen

3.1 Summary

In clinical research, the Cox proportional hazards model is widely used for analyzing survival data. One drawback of the method, however, is its assumption that covariate effects remain constant over time, and do not depart from proportionality. If this proportional hazards assumption is violated, the effect of the associated variable can instead be modeled as time-varying, and the most flexible way to model this effect is via non-parametric methods, which do not rely on further assumptions about the underlying shape of the time-varying effect. We compared five different non-parametric approaches for estimating time-varying regression coefficients in a Cox model. Four of these methods, namely Bernstein polynomials, B-splines, penalized splines, and restricted cubic splines, use basis polynomials to estimate $\beta(t)$ as a function of time. The fifth, local linear estimation, uses a kernel-weighted log partial likelihood evaluated across t to estimate $\beta(t)$. We evaluated the relative performance of these methods via extensive simulation studies, and then applied those with the best performance to a data set of 40,607 subjects with various types of tumors. We used the Cox model to assess the relationship between metformin exposure and overall survival time from cancer diagnosis after adjusting for demographic, disease, and drug information from electronic health records.

3.2 Introduction

The Cox proportional hazards model, one of the most popular methods for analyzing survival data, relates the hazard of an individual subject to their respective covariate set multiplicatively (Cox, 1972). While the model is straightforward to interpret via hazard ratios, it does so while assuming that the effects of all covariates remain constant over time. There are methods to handle departures from proportionality, such as via stratification on covariates that violate the PH assumption or partitioning the time axis (Therneau and Grambsch, 2000); however, stratification on any continuous variable requires discretization, and partitioning of the time axis requires the assumption of proportionality within each partition. Instead, the coefficient effect itself can be modeled as a function of time t, denoted $\beta(t)$, departing into non-proportional hazards models.

There exist multiple established methods of estimating time-varying coefficients within the Cox regression framework. Particularly useful are non-parametric methods, which offer flexibility in estimating $\beta(t)$ without having to make assumptions about the underlying relationship between time and the value of the coefficient (Chen et al., 2013). One such approach is via basis functions, which can approximate an unknown functional form of $\beta(t)$ by instead estimating a coefficient for each term in the basis function defined over t. A variety of basis functions have previously been applied to the area of time-varying coefficients, and we chose to consider four of them: Bernstein polynomials (Osman and Ghosh, 2012), B-splines (Chen et al., 2013; Sleeper and Harrington, 1990), penalized splines (Eilers and Marx, 1996), and restricted cubic splines (Austin et al., 2022). Each have been evaluated in terms of inferential validity within the Cox model, and each can be fit using existing software packages once the basis functions have been evaluated. However, each method can be tuned in performance through different means, such as the order of Bernstein polynomials, the number and location of knots for B-splines and restricted cubic splines, and the selection of a penalty term for penalized splines. The selection of each of these tuning parameters will impact the performance of the estimate of $\beta(t)$, particularly with respect to coverage, bias, and error (Perperoglou et al., 2019), though to what extent relative to others has not been previously examined.

Another existing method of estimating time-varying coefficients is local linear estimation, which uses a kernel smoothing technique to estimate $\beta(t)$ at pre-specified time points of interest. This method can be tuned by adjusting the bandwidth of the kernel estimator as well as modifying the distance between estimated time points t, with smaller intervals yielding more articulation in the final curve estimate but also higher computation time (Cai and Sun, 2003). Local linear estimation has been proposed as a useful method for visualizing and diagnosing time-varying effects of a covariate, but the need to maximize the log partial likelihood at every time point can be limiting when compared to basis methods, which can provide estimates at any time point t across the support of the basis once the basis coefficients are estimated. As a result, local linear estimation can be more computationally intensive and less flexible post-estimation. Moreover, its performance relative to other methods in ability to approximate an unknown $\beta(t)$ has not previously been evaluated.

In this work we compared the relative performance of the previously mentioned estimation methods via extensive simulation studies. We considered five different functional forms for the true $\beta(t)$, varying in shape and complexity, as well as both binary and continuous covariates on which we estimate the time-varying effect. Within each method we modeled over a range of tuning parameter settings, and both within and between methods we compared the ability to approximate $\beta(t)$, confidence interval width, confidence interval coverage probability, and mean square error. We also compared computation time of different methods. Based on these simulation results, we then applied the compared methods to evaluate the potential time-varying effect of metformin on cancer mortality (Xu et al., 2015; Wu et al., 2019).

3.3 Review of Methods

3.3.1 The Cox Model with Time-Varying Effects

In the usual proportional hazards survival model with right censoring, the available data is of the form $(Y, \Delta, \mathbf{X})$, where Y is the observed time defined as $Y = min(T, C)$, with T the event time and C the right censoring time, $\Delta = I(T \leq C)$, and $\mathbf{X}$ a vector of length p of covariates believed to relate to the outcome. The hazard at time t under this proportional assumption can then be denoted

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\beta^T\mathbf{X}) \tag{3.1}$$

where β is a vector of time-independent regression coefficients of length p and $\lambda_0(t)$ is an arbitrary baseline hazard function (Klein and Moeschberger, 2003). However, this model can be extended to non-proportional hazards by allowing the elements of β to vary with time (Hastie and Tibshirani, 1993). Throughout this manuscript, we consider the specific case where $\mathbf{X} = (X, \mathbf{Z})$, where X is a specific covariate whose effect we allow to vary over t while the remaining covariates $\mathbf{Z}$ are assumed to have time-independent effects, denoted γ, producing a hazard at time t of

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\beta(t)X + \gamma^T\mathbf{Z}) \ .$$

When fitting this model, approximation of $\beta(t)$ can be achieved in a number of ways, such as linear functions, splines, functional polynomials, and kernel smoothing (Austin et al., 2022; Chen et al., 2013; Cao et al., 2010; Cai and Sun, 2003; Zucker and Karr, 1990; Tian et al., 2005). In Section 3.3.2 we describe the five particular estimation methods whose performance we have chosen to compare.

3.3.2 Methods of Estimation

3.3.2.1 Bernstein Polynomials

Bernstein polynomials are linear combinations of Bernstein basis polynomials (Lorentz, 2013). For a given degree w, the corresponding Bernstein basis polynomials can be defined as

$$B_{l,w}(x) = \binom{w}{l} x^l (1-x)^{w-l}, \ \ l = 0,\ldots,w$$

where $\binom{w}{l}$ is a binomial coefficient. We can approximate $\beta(t)$ by defining it as a Bernstein polynomial of degree w, written

$$\beta(t) = \sum_{l=0}^{w} \alpha_l B_{l,w}(t)$$

where the coefficients α_l are called Bernstein coefficients. Previous work by Tenbusch (1997) demonstrated the performance of Bernstein polynomials in the regression setting, and Osman and Ghosh (2012) proved its validity in the estimation of time-varying effects for right-censored survival data, for both binary and continuous covariates.

3.3.2.2 B-splines

The B-spline basis is a type of spline base defined by its order w and number of internal knots K_n. It is a piece-wise polynomial with conditions ensuring continuity of both the function and its derivatives where the pieces join at each of the K_n knot locations (Sleeper and Harrington, 1990). To approximate $\beta(t)$ using B-splines, we assume that

$$\beta(t) = \sum_{l=1}^{d_n} \alpha_l B_l^w(t)$$

where $d_n = w + K_n$ is the total number of functions estimated (or $d_n = w + K_n + 1$ when including an intercept in the base) and α_l is the coefficient of the lth term in the B-spline approximation (Chen et al., 2013). A B-spline basis of order w with K_n internal knots defined over $t \in [0, \tau]$ will have a knot vector of $p_1 = \cdots p_w = 0 < p_{w+1} < \cdots < \tau = p_{w+K_n} = \cdots \; p_{2w+K_n}$. The spacing between knots in survival analysis is often based on quantiles of the observed event times under study, intending to include an equal number of events within each interval (Harrell et al., 2001). The individual basis function $B_l^w(t)$ can be defined via the Cox-de Boor recursion formula (De Boor, 1978) in polynomial pieces as follows:

$$B_l^0(t) := \begin{cases} 1 & \text{if } p_l \leq t \leq p_{l+1} \\ 0 & \text{otherwise} \end{cases}$$

$$B_l^w(t) := \frac{t - p_l}{p_{l+w} - p_l} B_l^{w-1}(t) + \frac{p_{l+w+1} - t}{p_{l+w+1} - p_{l+1}} B_{l+1}^{w-1}(t) \; .$$

The first component, $B_l^0(t)$, is a piecewise constant that is either 1 or 0 and indicates which knot span contains t. In the recursion portion, the first part, $\frac{t-p_l}{p_{l+w}-p_l}$, increases from 0 to 1 as t goes from p_l to p_{l+w}, while the second part, $\frac{p_{l+w+1}-t}{p_{l+w+1}-p_{l+1}}$, goes from one to zero as t goes from p_{l+1} to p_{l+w+1}. Outside of these ranges, the corresponding basis terms are zero.

3.3.2.3 Penalized Splines

Penalized splines, also referred to as p-splines, are a combination of the previously defined B-splines and penalization of the partial log-likelihood (Eilers and Marx, 1996). The goal of penalized splines is to prevent overfitting, particularly in the context of noisy data (Ruppert, 2002). To do this, $\beta(t) = \sum_{l=1}^{d_n} \alpha_l B_l^w(t)$ is estimated over d_n B-splines as previously defined in Section 3.3.2.2. However, estimates for the B-splines parameter set α are found by maximizing the penalized partial log-likelihood

(Malloy et al., 2009),

$$l^{\theta}(\alpha) = l(\alpha) - \theta \int [\beta''(t)]^2 dt$$

where $l(\alpha)$ is the partial log-likelihood for the overall model with respect to the spline coefficients α and θ, which controls how the curvature of $\beta(t)$ is penalized via its second derivative, determining the shape of the resulting estimate $\hat{\beta}(t)$. Different approaches for selecting the optimal smoothing parameter θ, as well as the number of knots k, have been proposed, such as via AIC, a corrected form of the AIC (Malloy et al., 2009; Hurvich et al., 1998), or different forms of cross-validation (Cao et al., 2010).

3.3.2.4 Restricted Cubic Splines

Cubic splines are a type of spline basis that use a set of cubic polynomials, each between two sequential knots, to represent a continuous variable. They are defined by the total number of knots k, which are placed across the range of data generally based on quantiles of, in this case, the unique observed event times (Harrell et al., 2001). A cubic polynomial is fit within each interval defined by the knots (for k knots there will be $k+1$ intervals in which to fit), and these polynomials are fit so that they meet smoothly with one another at the knots (Austin et al., 2022). The behavior of regular cubic splines can be poor in the tails of the curve (before the first and after the final knot, also called the boundary knots). This can cause issues when extrapolating the fit beyond the boundaries (Harrell et al., 2001), which can be addressed by instead fitting restricted cubic splines. In a restricted cubic spline, the tails of the function are forced to be linear. In doing so, the number of fitted cubic polynomials reduces to $k-1$, and the restricted cubic spline function used to estimate $\beta(t)$ can be defined as follows. We estimate

$$\beta(t) = \sum_{l=1}^{k-1} \alpha_l B_l(t)$$

where $B_1(t) = t$ and for knots $p_1, p_2, \ldots, p_k$, for $j = 1, \ldots, k-2$ we have

$$B_{j+1}(t) = (t-p_j)_+^3 - \frac{(t-p_{k-1})_+^3(p_k - p_j)}{(p_k - p_{k-1})} + \frac{(t-p_k)_+^3(p_{k-1} - p_j)}{(p_k - p_{k-1})}$$

where $(w)_+ = w$ when $w > 0$, and 0 otherwise (Harrell et al., 2001).

3.3.2.5 Local Linear Estimation

In contrast to the previous four methods, which all incorporated basis functions, time-varying coefficients can be estimated using a local partial likelihood technique, developed by Cai and Sun (2003). In this approach, $\beta(t)$ is estimated locally at pre-specified time points via calculation of the partial likelihood in a neighborhood around each time point. More specifically, for the hazard function defined in (3.1), the log partial likelihood across the interval $(0,\tau)$, where $\tau > 0$ is the maximum follow-up time, can be written

$$l(\mathbf{a}) = \sum_{q=1}^{n} \int_0^{\tau} \left[\mathbf{X}_q(s)^T \mathbf{a}(s) - \log \left\{ \sum_{i=1}^{n} Y_i(s) \exp(\mathbf{X}_i(s)^T \mathbf{a}(s)) \right\} \right] dN_q(s)$$

where $\mathbf{a}(t) = (a_1(t),\ldots,a_p(t))^T$ are the coefficients for the covariates $\mathbf{X}$, $N_i(t) = I(Y_i \le t, \delta_i = 1)$ is the counting process of observed failures for the ith individual, and $Y_i(t) = I(Y_i \ge t)$ is an indicator of the ‘at risk’ process. For this method, it is assumed that the coefficient function $\mathbf{a}(s)$ has a continuous second derivative within the neighborhood of t and that $\mathbf{X}(t)$ is a locally bounded predictable process. In this context, for s in a neighborhood of t and $j = 1,2,\ldots,p$, by Taylor’s expression,

$$a_j(s) \approx a_j(t) + a_j'(t)(s-t) \tag{3.2}$$

Define $\beta = (a_1(t),\ldots,a_p(t),a_1'(t),\ldots,a_p'(t))^T$ and $\tilde{\mathbf{X}}_i(u,u-t) = \mathbf{X}_i(u) \otimes (1,u-t)^T$ with $\otimes$ being the Kronecker product. Further, let $h = h_n > 0$ be a bandwidth parameter controlling the size of a local neighborhood and $K(\cdot)$ be a kernel function that smoothly down-weights the contribution of remote data points. Then, by the local linear model (3.2), incorporating localized weights, the local linear partial likelihood function is obtained by

$$l(\beta) = \sum_{q=1}^{n} \int_0^{\tau} K_h(u-t) \left[\tilde{\mathbf{X}}_q(u,u-t)^T \beta - \log \left\{ \sum_{i=1}^{n} Y_i(s) \exp(\tilde{\mathbf{X}}_i(u,u-t)^T \beta) \right\} \right] dN_q(u) \tag{3.3}$$

where $K_h(\cdot) = K(\cdot/h)/h$. Let $\hat{\beta}$ be the maximum likelihood estimate of (3.3) with respect to β. Then the local linear partial maximum likelihood estimate of $\mathbf{a}(t)$, $\hat{\mathbf{a}}(t)$, is the first p components of $\hat{\beta}$, while the last p components estimate the derivative of $\mathbf{a}(t)$. The first components of this vector is of interest for this manuscript, as our goal is estimation of the time-varying coefficient on X in particular. Further details, including a derivation of the estimation of the variance of $\hat{\mathbf{a}}(t)$, can be found in Cai and Sun (2003) and Tian et al. (2005).

3.4 Simulation Study

3.4.1 Data Simulation

The datasets for this work had 2,000 subjects per simulation with a hazard function for subject $i \in (1,2,\ldots 2000)$ defined as

$$\lambda(t|X_i,Z_i) = \lambda_0(t)\exp(\beta(t)X_i + \gamma Z_i)$$

where X_i and Z_i are either both binary or both continuous covariates with some level of correlation within subject. In the case where both are binary, we first sample $Z_i \sim Bin(0.5)$, and then sample X_i conditional on Z_i such that $X_i|Z_i = 0 \sim Bin(0.4)$, while $X_i|Z_i = 1 \sim Bin(0.6)$. When Z_i and X_i are both continuous, we first sample $Z_i \sim Unif(0,1)$ and then sample X_i as a function of Z_i such that $X_i|Z_i \sim 0.3Z_i + Unif(0,1)$. Across all simulations, $\lambda_0(t) = 0.2$ and $\gamma = 0.15$. Censoring times were generated from the exponential distribution with hazard function $\exp(0.1+0.1Z_i)$, truncated by 0.05 and the largest follow-up time of $\tau = 1.8$.

To evaluate performance of these methods across different shapes of $\beta(t)$, we considered five functional forms of differing complexities:

(a) $\beta(t) = 0.1$, a time-independent effect

(b) $\beta(t) = 0.1\log(t+1)$, a monotone increasing effect

(c) $\beta(t) = \log(0.7 + 2t\exp(-t^2))$, a non-monotonic effect with the highest hazard at the midpoint that changes direction gradually

(d) $\beta(t) = \log(1.2 - 0.9\sin(2t))$, a non-monotonic effect with the lowest hazard at the midpoint that changes direction sharply

(e) $\beta(t) = 0.3 - 0.5\cdot(\sin(2t) + \cos(2\sqrt{3}\,t))$, a non-monotonic effect with changing curvature at both the midpoint and tails

The shapes of these coefficients across $t \in (0, 1.8)$ can be found in Figure 3.1. Each of these values of $\beta(t)$, in conjunction with the above simulation settings, produce event rates between 12 and 15%. For each combination of $\beta(t)$ and data type (binary or continuous) (10 in total) we simulated 1000 data sets, resulting in 10,000 data sets simulated and fit by all methods.

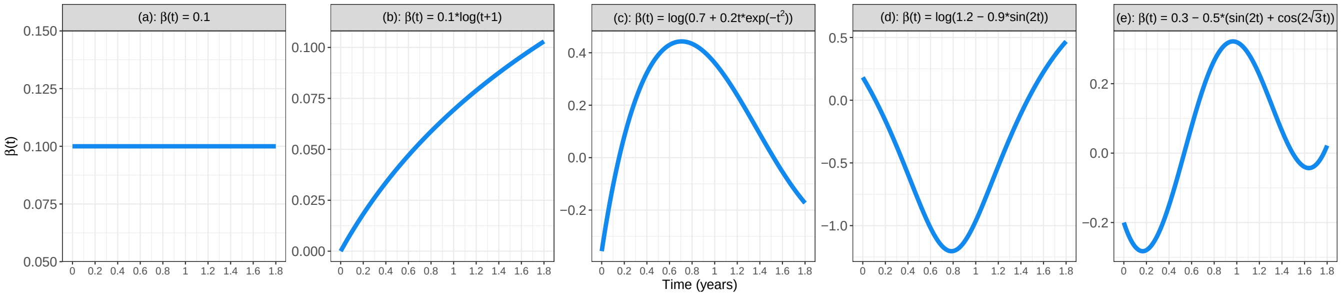

Figure 3.1: Considered functional forms of $\beta(t)$ by increasing complexity.

3.4.2 Tuning Parameter Selection by Estimation Method

As explained in Section 3.3.2, each of the considered estimation methods has its own tuning parameters to consider when fitting. For both B-splines and restricted cubic splines, the number and location of knots varied. We considered fits with $k = 3, 4, \ldots, 7$ knots, with placement of these knots chosen based on quantiles of the observed unique event times following the suggestion of Harrell et al. (2001), detailed in Table 3.1. All B-spline fits were fourth order, or third degree. For Bernstein polynomials, orders 3 through 6 were fit in each setting. For penalized splines, following the conclusions of Malloy et al. (2009), the corrected AIC (AICc), proposed in Hurvich et al. (1998) was used to identify the optimal smoothing parameter λ for each fit. Finally, for local linear estimation, bandwidths between 0.2 and 0.6 in intervals of 0.1 were considered (Liu et al., 2010; Cai and Sun, 2003), with the likelihood function evaluated at intervals of 0.05. To compare between different tuning settings per estimation method, each simulated data set was fit by all candidate tuning settings, and performance compared between them both at the data set level and at the summary level, as explained in Section 3.4.4.

Table 3.1: Quantile placement of knots for B-splines and restricted cubic splines based on number of knots k

k	Quantiles						
3			0.10	0.50	0.90		
4			0.05	0.35	0.65	0.95	
5		0.05	0.275	0.50	0.725	0.95	
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.50	0.6583	0.8167	0.975

3.4.3 Model Fitting Procedure

For all approaches using basis polynomials (Bernstein, B-spline, penalized spline, restricted cubic spline), model fitting was performed using the **coxph** function in the **survival** package version 3.1-8 in R version 3.6.0 (Therneau, 2015; R Core Team, 2020). The **tt** option was used to fit $\beta(t)$, multiplying X by the corresponding basis matrix over event times t. For Bernstein polynomials, the **Bernstein_basis** function from the **basefun** package version 1.1-3 was used to define the basis polynomials of order w over support $(0, \tau)$ (Hothorn, 2020). For B-splines, the basis matrix was generated using the **bs** function from the **splines** packages version 3.6.3, including an intercept term and set to the default of degree 3. Penalized splines were fit using the **pspline** function, part of the **survival** package. All penalized splines were of degree 3, and "optimal" degrees of freedom were chosen based on the corrected AICc. These optimal degrees of freedom also determined the selection of the penalty term θ as well as the number of splines in the basis. Finally, the restricted cubic spline basis polynomials were generated using the **rcspline.eval** function from the **Hmisc** package version 4.2-0 (Harrell Jr et al., 2020).

For all estimates of β generated via local linear estimation, the local linear partial likelihood function found in (3.3) was maximized using the **optim** function from the **stats** package version 4.0.5 with a max iteration limit of 10,000, optimized using the Nelder-Mead method (Nelder and Mead, 1965).

To obtain the variance of $\beta(t)$ at any particular time point t for all basis polynomial methods, we use the set of basis functions $\mathbf{B}(t)$ and the elements of the variance-covariance matrix Σ corresponding to α, denoted Σ_α, to get $\widehat{Var}(\beta(t)) = \mathbf{B}(t)\Sigma_\alpha \mathbf{B}(t)^T$

For local linear estimation, the variance of $\beta(t)$ is calculated at each time point t following Cai and Sun (2003) based on the second derivative of the local linear partial likelihood function.

3.4.4 Methods of Evaluation

Candidate estimation methods were first evaluated within methods, and then between methods. For each repetition of each simulation setting, method, and set of tuning parameters, the resulting estimate of $\beta(t)$ was evaluated for $t \in (0, 1.8)$. The median across simulations at each time point was then calculated, such that for a given time t we report $\tilde{\beta}(t)$ as the median of $\hat{\beta}(t)$, a vector of length 1000 where the jth element is $\hat{\beta}_j(t)$, the estimate from the jth simulation repetition, for $j \in 1, 2, \ldots, 1000$. This is repeated across t in increments of 0.01 for all methods except local linear estimation, which was evaluated in increments of 0.05. Similarly, 95% confidence interval bounds around each $\hat{\beta}_j(t)$ are calculated as $\hat{\beta}_j(t) \pm 1.96 * SE(\hat{\beta}_j(t))$ for each candidate method, based on the fitted variance of $\beta(t)$ described in Section 3.4.3. The medians of the lower and upper bounds of these confidence intervals are then calculated and reported for each value of t as the "Fitted SE" confidence intervals. For comparison, we also calculate 95% empirical confidence intervals around $\tilde{\beta}(t)$ based on the standard deviation of $\hat{\beta}(t)$ at each time point t, calculated as $\tilde{\beta}(t) \pm 1.96 * SD(\hat{\beta}(t))$. These confidence intervals are reported as "Empirical SD." Alignment of the two types of confidence interval bounds reflects the ability of the variance estimator to capture the empirical variance in $\beta(t)$.

In addition to these summary measures, relative performance was evaluated based on coverage, mean square error, AIC or log-likelihood, and computational time. Coverage at time point t is calculated as the percentage of simulation repetitions whose lower and upper bounds contained the true value of $\beta(t)$, out of 1000 repetitions. Mean square error is calculated per time point as $MSE(t) = \frac{1}{n}\sum_{j=1}^{n}(\beta(t) - \hat{\beta}_j(t))^2$.

While the above metrics summarize the performance specific to each tuning parameter selection, we additionally wanted a way to compare between tuning parameters to find the "optimal" performance of an estimation method, so that this approach could be compared to the other methods. To do so, we calculated model AIC for Bernstein polynomials, B-splines, and restricted cubic splines. For each simulated data set, we compared AIC across tuning parameters and considered the lowest AIC to be the optimal fit for that data. Once the optimal fits were selected for each of the 1000 simulated data sets, their performance was summarized as previously described in terms of the median curve, 95% confidence intervals, coverage, and mean square error. For penalized splines, as the optimal smoothing parameter λ was already selected during model fitting via AICc, the resulting summarized curve was considered to be the optimal fit.

Finally, for local linear estimation, we first calculated the log likelihood at each time point for each data set and selection of tuning parameters. For a given simulation j with bandwidth m, we calculated the mean log likelihood across t as $\overline{ll(t)}_{j,m} = \frac{1}{n_j}\sum_{v=0}^{t} \log(L(\hat{\beta}(v)_{j,m}))$, where $n_j = 37$ is the total number of time points estimated for each fit. The optimal local linear estimation tuning set for simulation j was then chosen as the bandwidth m that produced the largest $\overline{ll(t)}_{j,m}$. Similarly to the above, the optimal fits for each of the 1000 simulated data sets were then summarized alongside the tuning-specific summaries. These procedures were intended to reflect model selection when fitting to real data and attempting to optimize tuning parameters to produce either the lowest AIC or the highest log-likelihood.

In addition to the ability of each method to approximate the true $\beta(t)$ curve, we were also interested in the computational efficiency of each estimation method. All simulations were executed using the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) cluster on an Intel Sandy Bridge architecture processor with 2 GB of system memory. Computation time per fit j was recorded and compared across tuning parameters and methods of estimation.

3.5 Simulation Results

3.5.1 Within Estimation Method

We begin with an assessment of the relative performance of different tuning parameters within each estimation method of interest. The estimated median curves, coverage probabilities, and mean square error for each method can be found in Figures 3.2 to 3.6, grouped by method. For all figures, panels are arranged by curve shape (a) to (e) left to right and by data type top to bottom, with binary X and Z on top and continuous X and Z on the bottom.

In Figure 3.2 we have Bernstein polynomial fits of orders 3 through 6 plotted alongside the “optimal” curve as defined in Section 3.4.4. Looking first at the estimates of $\beta(t)$ in Panel **A**, we find that for curve shapes (a) and (b), there is no real difference between the orders in terms of ability to estimate the true curve. Moving to the right, however, as the curves become more complex it appears that increasing the order improves the ability of the Bernstein polynomial basis to approximate the shape of $\beta(t)$. This is particularly visible for curve (d) in the trough of the curve around $t = 0.75$, and in curve (e) in the extreme tails of the curve. As the order increases, the width of the 95% confidence intervals increase as well. In

Panel **B**, coverage for curves (a) through (c) is close to the nominal 95% level across t. For curve (d), coverage is low in the binary data case for orders 3 and 4 as well as the optimal curve, particularly around the curve trough. Similarly, coverage is low for curve (e) in the binary case around the left tail, as the lower order methods were unable to capture the changing curvature there. Finally, looking at Panel **C**, we see that mean square error increases with increasing order across all curve shapes, particularly in the end of the time range from $t = 1.5$ onward. For both panels **A** and **C**, the behavior of the "optimal" curve is most similar to the curves of order 5.

Looking next to Figure 3.3, we summarize B-spline polynomial fits of degree 3 with 3 - 7 knots. In Panel **A**, we find that each knot specification performs well in estimating the shape of the curve for (a) through (c), while in curve (d) it is evident that fits with higher numbers of knots are better able to capture the trough of the curve. However, in curve (e) the different specifications are all able to capture the extreme tails fairly well. Confidence interval widths increase with increasing number of knots. In Panel **B** we find that all curves maintain at or slightly above nominal coverage for both types of data, with coverage above 95% occurring in the extreme tails of fits where data is more sparse. In Panel **C**, mean square error increases with number of knots, but this is more pronounced with binary data than with continuous. The performance of the optimal curve is closely aligned with a 5 knot fit in both panels **A** and **C**.

The penalized spline fit can be found in Figure 3.4. As each of the 1,000 repetitions was fit with an optimal number of degrees of freedom, there is only one summary curve for these simulations. In Panel **A**, curve (a) misses slightly in either tail, while curve (b) is more accurate. For curves (c) and (d), the penalized spline struggles to capture both the tails and the peak of the curve. It is similar in (e), though the extreme behavior of the tails coupled with the peak in the middle, when being fit with the penalized spline, results in a very smooth median curve with little articulation. Confidence intervals around the estimate widen across time. In Panel **B**, we see lower coverage at early values of t for curve (a), nominal coverage for curve (b), and for curves (c) through (e) we have lower coverage surrounding most points of articulation along the curve, likely due to the overly smooth fits produced by the penalized spline. Mean square error in Panel **C** increases across time, particularly for curves (c) and (d).

In Figure 3.5 we have the restricted cubic spline fits with 3 - 7 knots. At only 3 knots, the median curves in Panel **A** are unable to capture any curvature of the true $\beta(t)$ for curves (c) through (e), but all numbers of knots perform alright for (a) and (b). With increasing knots, the peaks and troughs of curves (c) through (e) can be captured, but the linear tail restrictions hamper the ability of the curve to approximate the tail behavior in (e). Confidence interval widths again increase with increasing number of knots. In Panel **B**, coverage is low at low numbers of knots, especially for 3 knots in curves (c) through (e) around points of articulation. Coverage is also low for low values of t for curves (c) and (d), due to the linear restriction through this range. Mean square error in Panel **C** increases with the number of knots, but is also high when there are 3 knots in curve (d) at the trough. The optimal curve behavior for restricted cubic splines aligns closely with orders 4 and 5 in terms of curve shape, but its coverage is very low for (c) through (e).

Finally, in Figure 3.6, the local linear estimation fits are summarized. In Panel **A** we see that narrow bandwidths are the most able to capture the articulation in curves (c) and (d), but their confidence intervals are also the widest. These fits particularly struggle to capture the trough of curve (d), as well as the right tail of curve (e). In Panel **B**, coverage is nominal through the majority of time for curves (a) through (c), but does drop below nominal in the highest values of t. For curve (d), the widest bandwidths have very low coverage in the trough, and all bandwidths fall below nominal at high t, which is also evident in curve (e). Finally, for mean square error in Panel **C**, the narrowest bandwidths suffer from the highest MSE. The optimal curve performs poorly in terms of confidence interval width and mean square error. However, its coverage is somewhere between a bandwidth of 0.2 and 0.3.

3.5.2 Between Estimation Methods

The best fit from each method, chosen as the tuning parameter selection that most closely mimicked the "optimal" fit in each case, is presented in Figure 3.7. When compared against other approaches, it is clear that the penalized spline fits are unable to capture $\beta(t)$ to a useful approximation, and their poor coverage reduces their validity in inference. Similarly, the restricted cubic spline with 5 knots suffers from poor coverage at early values of t for curve (c), and although its MSE is low it fails to capture the articulation in several of the curves. The Bernstein polynomial fits and local linear fits, at their best, perform comparably in both MSE and curve estimation, though their coverage suffers slightly at high values of t. However, both of these fit approaches showed high variability in performance depending on the tuning parameters chosen. Finally, the B-spline fit at its best maintains nominal coverage probability, and its confidence intervals are comparable in width to Bernstein polynomials and local linear estimation. It manages to capture the curvature of $\beta(t)$ across all curves, including in the tails of (e) and the trough of (d). While its MSE is the highest among the methods, it is also the only method that was able to maintain nominal coverage regardless of tuning parameters.

Beyond curve estimation, coverage, and MSE, we also looked at the computational efficiency of each method by comparing computation time per model fit. Restricted cubic spline fits were the fastest at 4-6 seconds per fit, depending on data type and curve shape. Both Bernstein polynomials and B-splines averaged 7-10 seconds per fit, with penalized the slowest among the basis function fits at 12-15 seconds. The longest fitting time by far, however, went to local linear estimation at 6 - 20 minutes per fit depending on curve complexity and data type. This was not unexpected given that the partial likelihood is evaluated at each time point of interest, and computation time reasonably scales with the interval of evaluation. These simulations estimated at intervals of 0.05, but modification of that interval width would dictate computation time.

3.6 Real-World Application: Metformin in Cancer Patients

3.6.1 Data Source and Methods

As an illustration, we built on previous work by Xu et al. (2015) and Wu et al. (2019) on drug repurposing in cancer patients. Their work found an association between metformin exposure and cancer mortality among patients with all tumor types, including breast, lung, and colorectal tumors. We extended on these findings by allowing the effect of metformin to vary across time. The data for this work comes from a retrospective cohort study, running from January 1, 1995 to December 31, 2010 using Vanderbilt University Medical Center (VUMC) electronic health records (EHRs) sourced from the Synthetic Derivative (SD), which is a comprehensive and de-identified version of the VUMC EHR database (Roden et al., 2008). We identified 40,607 total subjects for consideration across all tumor types, with 2,951 among these having documented metformin exposure in the SD. Of these identified subjects, there were 3,963 with breast tumors, 3,357 with lung tumors, and 2,542 with colorectal tumors.

Our modeling approach followed the procedure outlined by Wu et al. (2019), beginning with pre-specified study covariates of age, biologic sex, race, tumor type, and tumor stage. Also under consideration were diseases with International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) codes, grouped into 1,637 phenome-wide association study (PheWAS) phenotypes (Denny et al., 2010), and information on 1,346 normalized medications, including metformin. To reduce the dimensionality of candidate covariates, variable screening using a univariable Cox proportional hazards model was conducted for each phenotype covariate, and only those with $p < 0.3$ in the univariable setting were kept in the final model. After screening, this resulted in 1,279 disease codes for inclusion in the final model. Outcome of interest was overall survival time from cancer diagnosis to death (event) or last medical record date in the EHR (censored), rounded to the nearest month. Outcomes were censored at a maximum follow-up time of five years. Exposure to metformin was modeled as a binary variable and fit with a time-varying coefficient effect following the procedure outlined in Section 3.4.3. Due to the extensive number of model covariates (2,413 total after univariable screening), local linear estimation was not feasible on this dataset due to its long computation time ($>$ 24 hours per time point evaluated). Additionally, penalized splines were not fit to this data due to their poor coverage in the simulation scenarios considered previously. The remaining three candidate methods were fit to all tumors types as well as individually to breast, lung, and colorectal tumors. Fits were based on the optimal tuning parameters previously discussed (5 knots for B-splines and restricted cubic splines, order 5 for Bernstein polynomials). For comparison, a model with a time-independent effect on metformin was also fit for each grouping.

3.6.2 Results

The estimated $\beta(t)$ curves for this analysis can be found in Figures 3.8 and 3.9. Looking first at Figure 3.8, we consider the performance of B-splines fit to all cancers, breast cancer, colorectal cancer, and lung cancer, and compare against a time-independent fit that assumes $\beta(t) = \beta$ for all t. Both colorectal and lung cancers did not show evidence suggesting a time-varying effect of metformin exposure, with estimates of $\beta(t)$ showing similar trends to β alone and particularly no evidence suggesting a bi-directional effect of exposure that would switch from protective to harmful or vice versa. In breast cancer, both the time-varying and time-independent effects contain 0 in their confidence intervals, and there seems to be no real directional trend in the estimate of $\beta(t)$ when comparing to β. Among all cancers, however, there is an upward trend in $\beta(t)$ that switches from protective ($\beta(t) < 0$) to harmful ($\beta(t) > 0$) as t increases. While the confidence intervals still contain 0, and overlap with the time-independent estimate, there is enough of a trend in this fit to suggest that a time-varying effect of $\beta(t)$ is plausible and justifies future investigation. In Figure 3.9, where all plausible methods are compared, we find similarly that the effect of metformin when modeled for all cancers shows a suggestion of a time-varying effect, with a similar increasing trend modeled by both Bernstein polynomials and restricted cubic splines compared to the B-splines fit. Fits for individual cancers also show similar shapes, though the B-spline fit shows the most variability among the three due to a higher level of flexibility in the fit. However, as it was shown that Bernstein polynomials and restricted cubic splines were more prone to low coverage, it is possible that the lower variability in these fits does not indicate a better overall fit. Either way, the findings pertaining to each cancer are consistent across the considered methods, and all support further investigation into a time-varying effect of metformin on cancer survival. Of note is that, consistent with simulations, each of the three methods applied to this data had comparable computational times at 94 - 96 total hours to reach convergence.

3.7 Discussion

We assessed the relative performance of five different methods of estimating time-varying coefficient effects in the Cox regression model. Within each method, we found that increasing the flexibility allowed by the tuning parameters (number of knots, order of the fit, size of the bandwidth) resulted in fits that better articulated the true shape of $\beta(t)$, but at the expense of increased mean square error and less efficient estimators. The penalized spline struggled even when allowed to fit to an optimal number of degrees of freedom, as the estimate could not articulate to the most extreme portions of any $\beta(t)$ functions considered. This may be due to the moderately low event rate under consideration (12-15%), as previous works have shown the selection of optimal degrees of freedom may be far from the "real" optimum in small data sets (Verweij and van Houwelingen, 1995). Similarly, the linear restriction in the tails of the restricted cubic spline fit greatly reduced its performance in the case of a $\beta(t)$ with curvature in the tails, such as curve (e), particularly with a low number of knots. Previously, it has been suggested that restricted cubic spline fits are can be particularly hampered by knot placement when the positioning is non-uniform over the covariate space (Durrleman and Simon, 1989); however, in cases with a large amount of censoring, placing the knots based on quantiles of observation times will inevitably lead to non-uniform spacing along t.

The performance of Bernstein polynomials was heavily dependent on the order of the polynomial being fit, with low orders suffering from low coverage throughout any points of articulation in $\beta(t)$. While higher order fits were better able to capture the true curve shape, they suffered from higher MSE particularly in the tails of the fit. There has been work to suggest that in the case of a small sample, the choice for the order of Bernstein polynomials should be around $m = \sqrt{n}$ (Osman and Ghosh, 2012). For larger sample sizes, it has been suggested to use AIC or BIC (Kooperberg et al., 1995), which would be reflected in our "optimal" curve, most closely in alignment to an order of 5. Local linear estimation, similarly, had performance that relied predominantly on the selection of the bandwidth parameter, h. At small values, the model achieved nominal coverage but suffered from high MSE, particularly at higher values of t. At large values of h, however, coverage suffered through points of articulation such as the trough in curve (d) and tails in curve (e). The "optimal" curve in terms of the average value of the partial log likelihood had notably high MSE and wide confidence intervals. Liu et al. (2010) previously found that the selection of a bandwidth parameter $h = O(n^v)$, with $1/5 < v < 1$, would be optimal to obtain an unbiased estimator. In our simulations, this would correspond to a bandwidth $.0005 < h < 0.22$, which as evidenced by 3.6 would lead to nominal coverage but some of the highest MSE of all simulations. It is worth noting that Cai et al. suggest the use of local linear estimation as a diagnostic tool to uncover departures from the proportional hazards model, but they do not claim it to be optimal from an inferential perspective (Cai and Sun, 2003).

Among all methods considered, B-splines were the only ones to maintain nominal coverage regardless of the number of knots included in the basis. The widths of the confidence intervals for different knot specifications did not vary to the degree of other methods, and MSE was comparable between knot specifications through most of the range of t. The flexibility of B-splines has been touted in previous works, particularly for their lack of sensitivity to knot placement, and ability to capture curve shapes with as few as four or five knots regardless of complexity (Sleeper and Harrington, 1990; Chen et al., 2013; Hastie et al., 1992). While the MSE for B-splines was higher than other methods in the right tails for binary fits, it was the only method among those compared that was able to maintain nominal coverage in the binary case regardless of tuning parameters. Additionally, it was among the least computationally intensive at 7-10 seconds per fit for a dataset of size $n = 2000$. For these reasons, among the methods tested it is reasonable to conclude that the use of B-splines in estimating time-varying effects in the Cox model provides a balance of flexibility, efficiency, and analytical validity that set it apart from the others, at least based on empirical evidence.

Our application to metformin exposure found similar shapes in the estimate of $\beta(t)$ across fits using B-splines, Bernstein polynomials, and restricted cubic splines. B-splines showed the highest degree of flexibility in the fit, and had comparable computation times compared to the other methods. The fit to data from all cancer types suggested a time-varying effect of metformin exposure on cancer survival up to five years, with the effect transitioning from protective to harmful around 3.5 years. The effect of metformin on cancer survival (Rizos and Elisaf, 2013; Xu et al., 2015; Wu et al., 2019) has commonly been concluded as protective, however these findings suggest that this protection may only persist for a few years, supporting further research into its long-term effects. Of note is that our data for this application did not include length of exposure to metformin or dosing, and having this additional longitudinal information would improve the accuracy of the estimate of $\beta(t)$.

There are a few limitations to this work. First, we only consider simulations and examples with one potential time-varying effect of interest. In practice, there may be multiple covariates with suspected time-varying effects, which would require additional computation time in the case of all basis functions but not necessarily for local linear estimation, which estimates all coefficients with the potential to vary by time. Additionally, in most cases we only consider a narrow range of tuning parameters for each method. While we did find diminishing returns in most examples as the flexibility increased, in practice it may be necessary to go beyond the displayed scope of parameters to find the optimal fit for a particular set of data. With increasing flexibility, particularly with B-splines, we observed high variability in the tails of the estimated $\beta(t)$. There have been several methods suggested for dealing with this variability, including correcting skew in the related covariate (Sleeper and Harrington, 1990). While we do not implement methods to correct for high tail variability, it is an area of potential future work. Finally, the results of this work are empirical and based on summarized simulations. A possible extension would be to prove the asymptotic behavior of each method to confirm which would come out as optimal, though this would likely depend on context and other factors beyond simply the estimating method at hand. Nevertheless, we believe these results provide support for the use of B-splines in particular in estimating time-varying effects in the Cox regression model, particularly as a potential first-line approach to accommodating suspected non-proportional hazards.

Acknowledgments

This work was conducted in part using the resources of the ACCRE cluster at Vanderbilt University, Nashville, TN. The authors would also like to thank Dr. Jeremy Warner and Dr. Hua Xu for their guidance on our real-world application. Data for the real-world application was obtained through NIH grant U24CA194215.

Funding

- E.A.S.W. was supported by NIH grant U24CA194215.
- R.T. was supported by NIH grant R01AI131771 and the VUMC Biostatistics Faculty Development Award.
- Q.C. was supported by NIH grants U24CA194215 and R01CA188214.

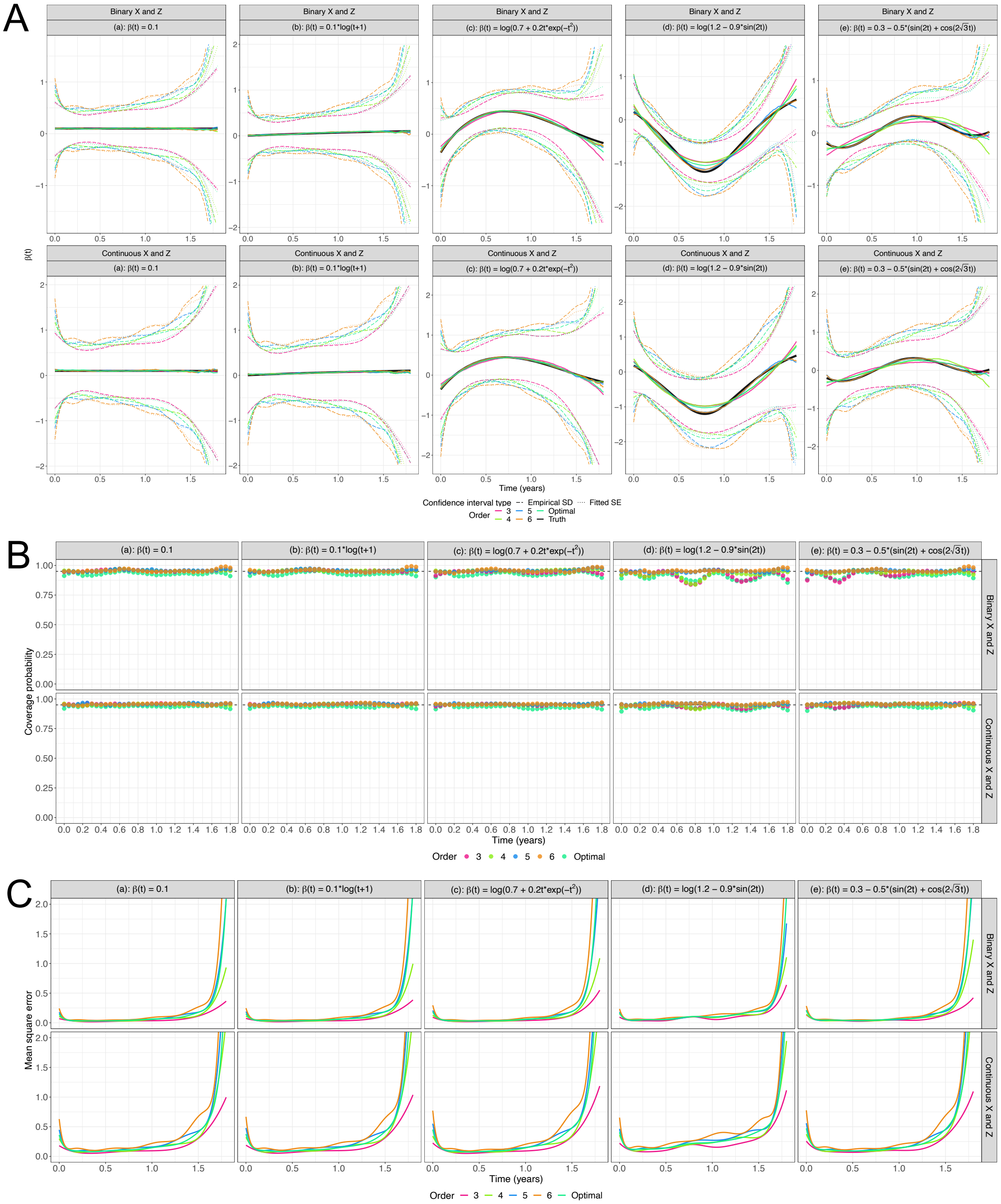

Figure 3.2: Simulation results for Bernstein polynomials, colored by order of fit. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability (C) Mean square error.

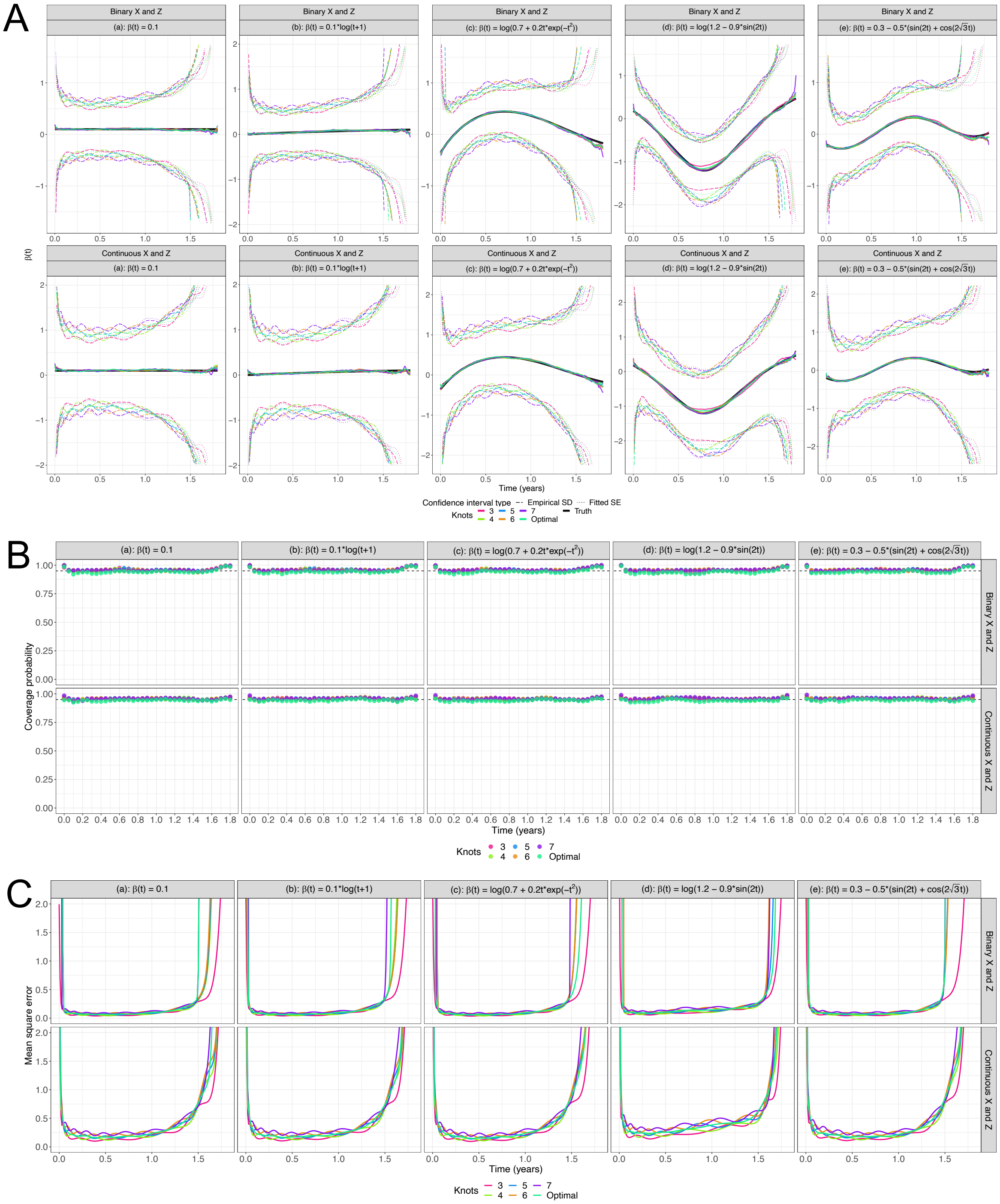

Figure 3.3: Simulation results for B-splines, colored by number of knots. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability (C) Mean square error.

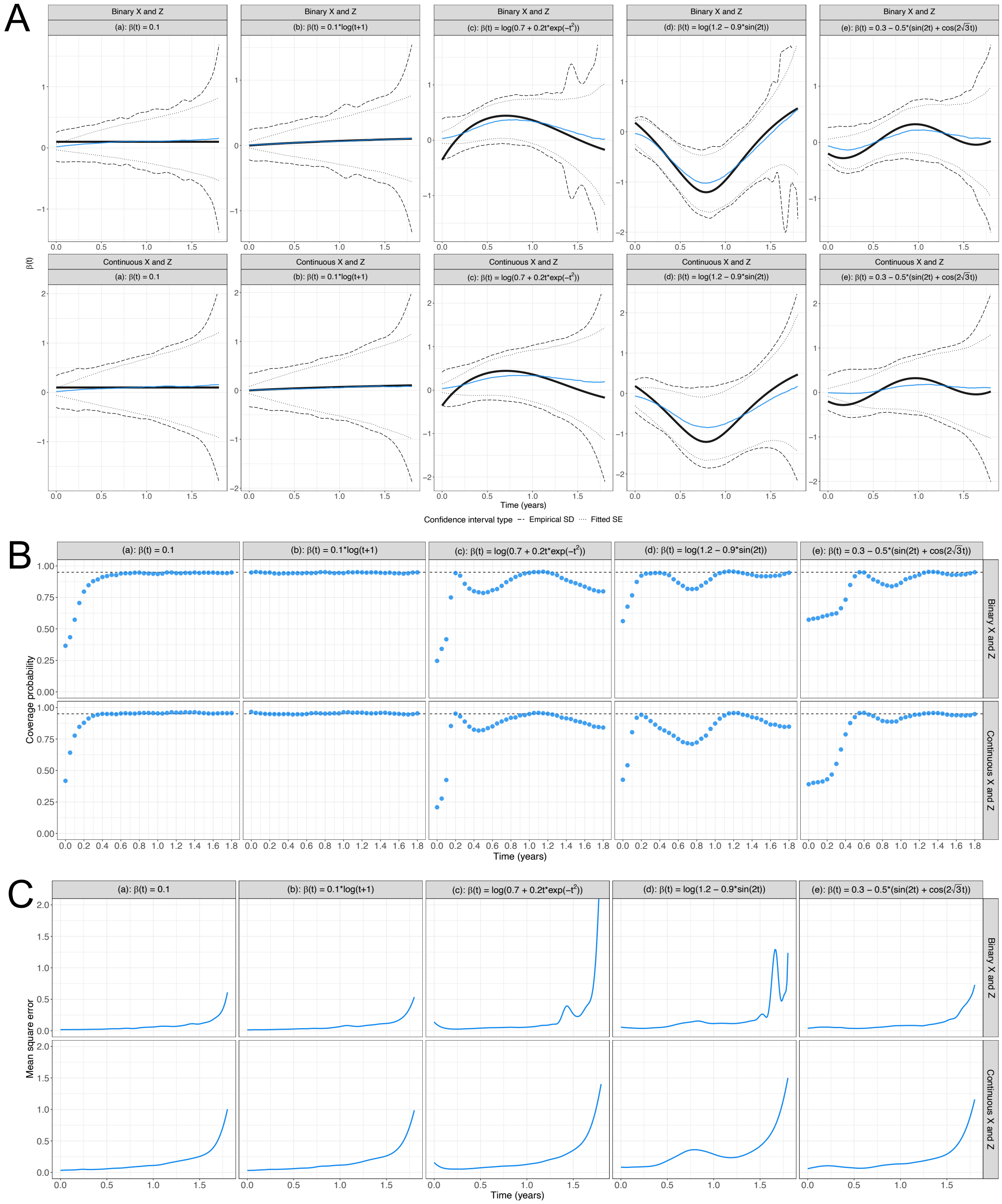

Figure 3.4: Simulation results for penalized splines. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability (C) Mean square error.

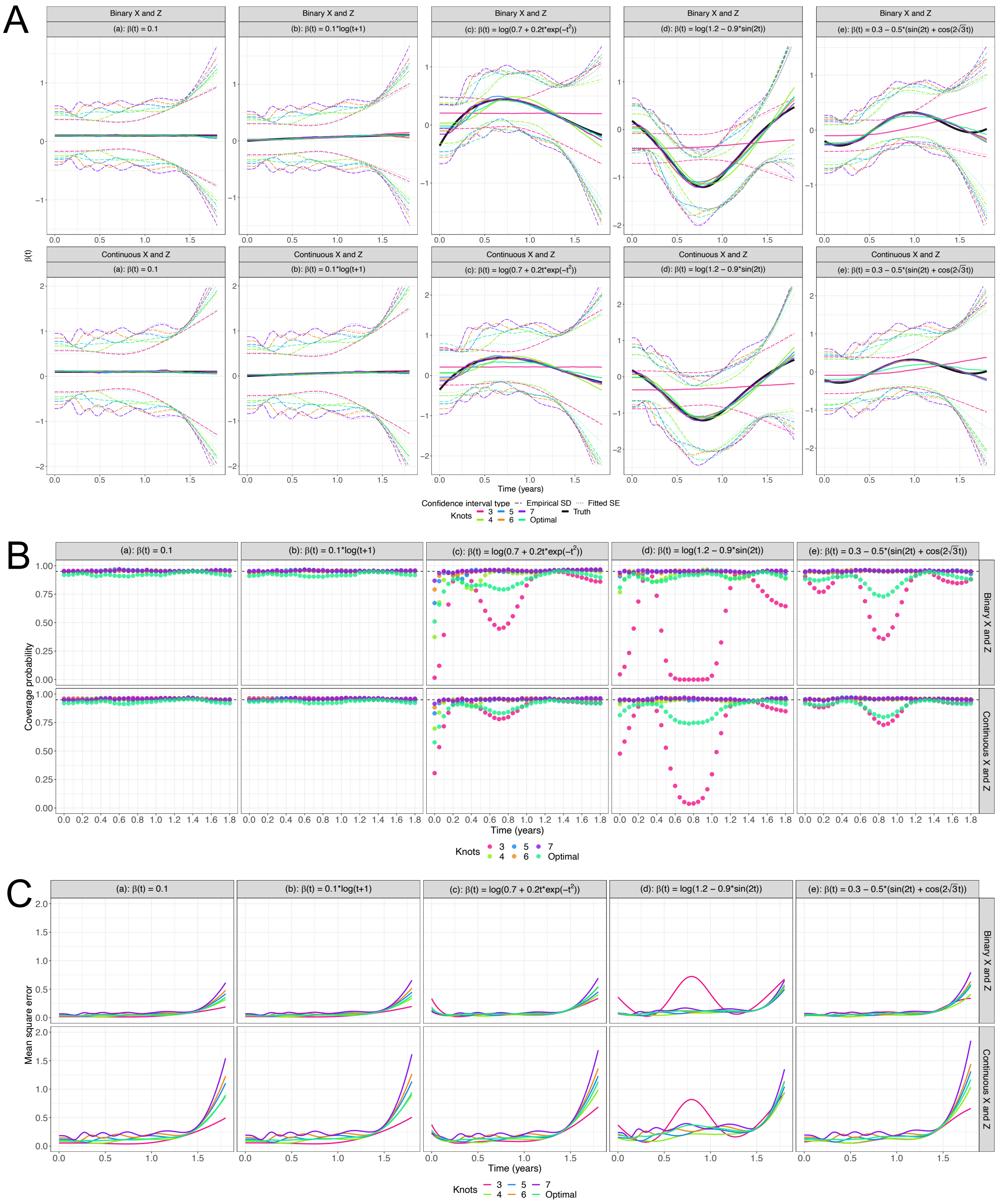

Figure 3.5: Simulation results for restricted cubic splines, colored by number of knots. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability (C) Mean square error.

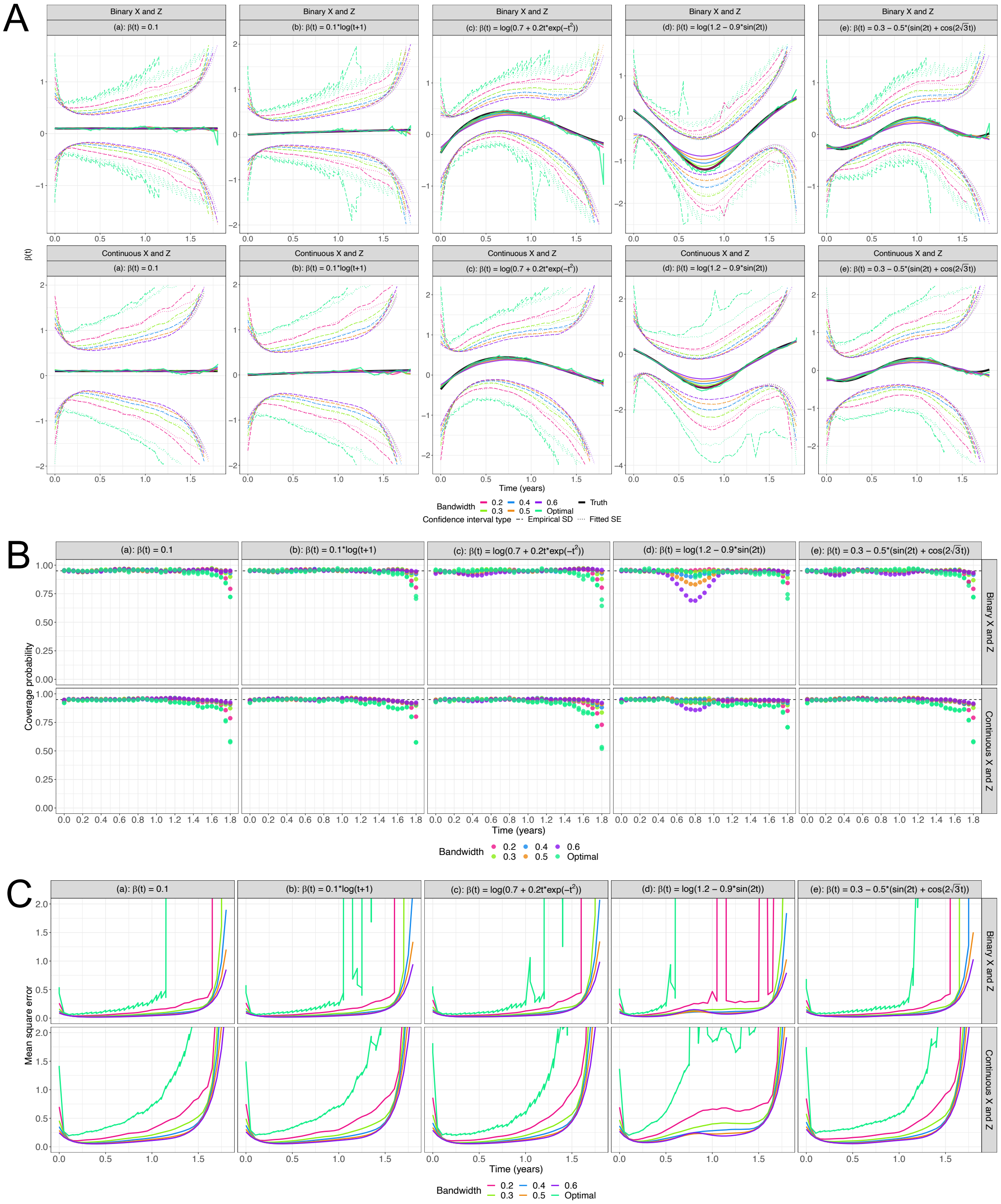

Figure 3.6: Simulation results for local linear estimation, colored by bandwidth. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability (C) Mean square error.

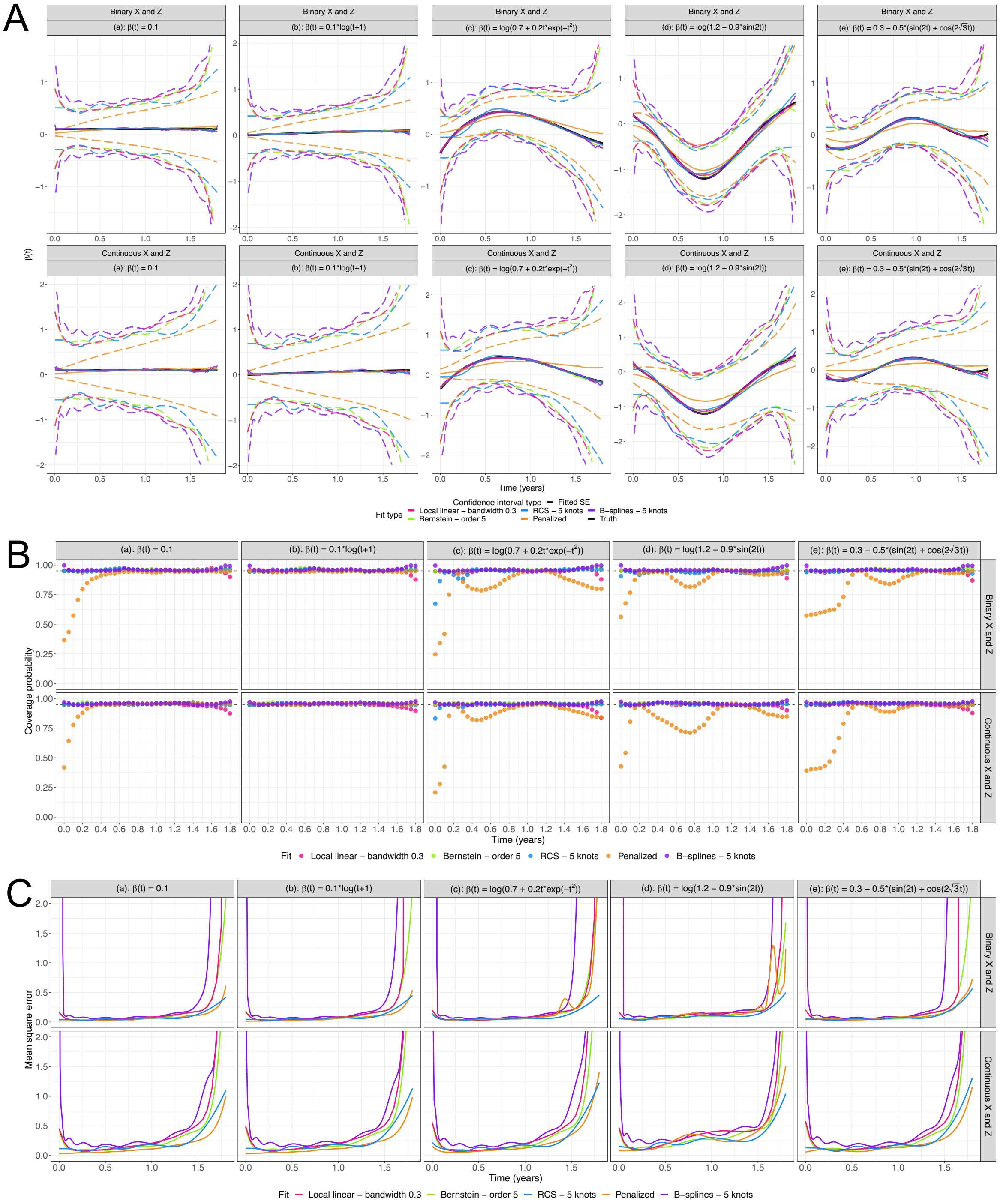

Figure 3.7: Simulation results from best fit of each method, colored by fit. (A) Estimated median curves and 95% confidence intervals (B) Coverage probability (C) Mean square error.

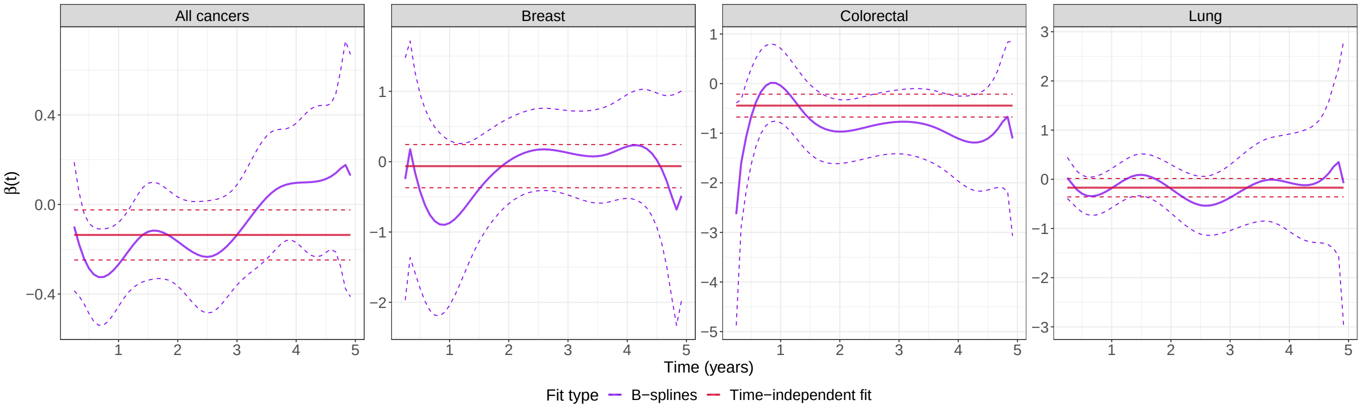

Figure 3.8: Estimated $\beta(t)$ curves for B-splines fit to metformin exposure data compared to a time-independent estimate of β.

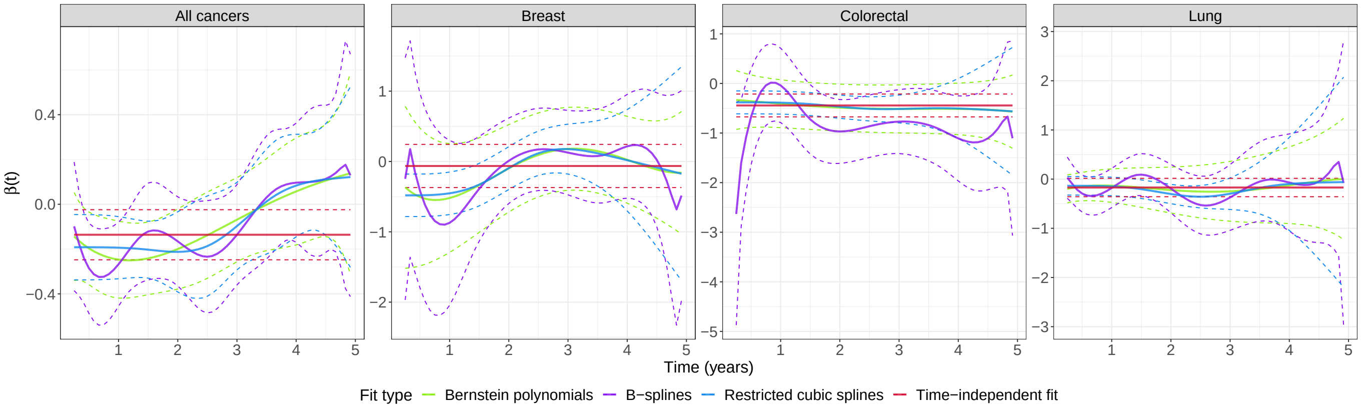

Figure 3.9: Estimated $\beta(t)$ curves for candidate methods fit to metformin exposure data compared to a time-independent estimate of β.

CHAPTER 4

Efficient Estimation of the Cox Model with Time-Varying Effects Under Two-Phase Designs

Elizabeth A.S. Westerberg, Ran Tao, and Qingxia Chen

4.1 Summary

Two-phase study designs are often used in large epidemiological or clinical studies where certain covariates are too expensive to be collected on all participants. In the case of time-to-event outcomes, expensive covariates are often collected on all cases along with a subset of the controls, sampled in different ways. The Cox proportional hazards model is a common model of choice for time-to-event outcomes, and previous work using sieve maximum likelihood estimators has developed an efficient semiparametric method for estimating the Cox model under two-phase designs when covariate effects are constant by approximating the conditional density functions of expensive covariates given available inexpensive covariates, using B-spline sieves. However, we are interested in extending this method to the case of time-dependent effects, as the long follow-up times often seen in large studies with rare events lend themselves to violations of the proportional hazards assumption. We propose an extension of this method that estimates time-varying effects for expensive covariates using B-splines. We establish its performance via extensive simulation studies and compare its performance to existing methods for estimating time-varying effects. Finally, we demonstrate our method on data from a large cohort study, looking at the association between oxidative stress and colorectal cancer incidence.

4.2 Introduction

In large epidemiological or clinical studies with a time-to-event outcome, the outcome of interest (time to diagnosis, time to death after diagnosis, length of progression-free survival, etc.) is often known for all subjects under study, as well as the majority of the demographic or inexpensive information that researchers believe may be associated with the event. However, the covariates of main interest are often values that involve genotyping, lab values, or imaging data, which are expensive as well as time-intensive or impossible to procure for all subjects. Oftentimes, studies deal with this difficulty by running what are referred to as "two-phase" studies, wherein the collection of inexpensive information is performed for all subjects, referred to as Phase I. In Phase II, the expensive covariates of main interest are measured for only a subset of the total population (Tao et al., 2020). Methods for sampling this second phase vary depending on the type of study. In survival analysis with time-to-event outcomes, common sampling schemes include the case-cohort and nested case control design, where in both all cases are included in Phase II. In the case-cohort design, a random sample of controls is selected from the total population for inclusion in Phase II (Prentice, 1986), while in the nested case control design controls are sampled for each case based on the risk set at the event time of each case, sampled in varying ratios of controls to cases depending on the study (Liddell et al., 1977).

While two-phase studies make data collection more feasible, they also require modifications to analysis techniques to account for the sampling method being used. One common analytical approach to time-to-event data in general is the Cox proportional hazards model, which links the hazard of a subject to their respective covariate set multiplicatively (Cox, 1972). It does so by assuming that the effects of model covariates remain constant over time, referred to as the proportional hazards assumption. In large data sets however, particularly those with rare events or long follow-up times, this assumption can often be violated, necessitating models that can handle departures from proportionality. As mentioned in Chapter 3, methods such as stratifying on covariates that violate the proportional hazards assumption or partitioning the time axis and estimating effects within partitions have been developed (Therneau and Grambsch, 2000). However, these are less ideal as they require discretization of stratifying variables or the assumption of proportionality within time partitions. It is preferable to model the time-varying coefficient, denoted $\beta(t)$, flexibly as a function of time. In particular, $\beta(t)$ can be modeled non-parametrically using B-splines, avoiding making assumptions about the underlying relationship between time and the value of the coefficient (Chen et al., 2013; Sleeper and Harrington, 1990). They can also be modeled via local linear estimation, where $\beta(t)$ is estimated at pre-selected time points of interest via kernel smoothing Cai and Sun (2003). Both of these approaches work for estimating time-varying effects, as evaluated in Chapter 3; however, as mentioned, they require additional modifications when dealing with two-phase designs.

Modifications can be sampling specific, such as the extension of local linear estimation methods to nested case control data by limiting the risk set to those sampled for a given case (Liu et al., 2010), or flexible to work with different types of sampling like inverse probability weighted models (Støer and Samuelsen, 2012), which incorporate different weights depending on the sampling type. Inverse probability weighting can be used in conjunction with B-splines for estimating time-varying effects, making it an option for a number of sampling designs. However, both of these methods drop individuals from Phase I who did not have the expensive covariate sampled in Phase II, resulting in a loss of information and reduction in total sample size. There exist methods that do incorporate Phase I data, such as the likelihood-based approach of Saarela et al. (2008), but these methods often require parametric assumptions about the association between the inexpensive and expensive covariates.

Existing work by Tao et al. (2017) developed efficient semi-parametric methods for several regression models under two-phase sampling, including for the Cox proportional hazards model. This method uses a B-spline sieve approximation to estimate conditional density functions for the expensive covariate of interest given the inexpensive covariates of choice, both discrete and continuous. In doing so, it is able to incorporate all Phase I information, including from those who were not selected for measurement of the expensive covariate in Phase II. We sought to extend this model to the case where the expensive covariate has a time-varying coefficient, building on the existing expectation-maximization algorithm to maximize the sieve likelihood. This approach combines the B-spline estimation method for time-varying coefficients as outlined in Chapter 3 and extends it to the two-phase case.

In this work, we begin by outlining an example use case of the Cox model with time-varying effects in a two-phase setting, with inspiration from a large cohort study. Next, we detail our method for efficient two-phase estimation of the Cox model with time-varying effects, detailing the B-spline sieves and expectation-maximization algorithm for coefficient estimation as well as a profile likelihood approach for estimating standard errors. We then compare the performance of our method against two previously established two-phase methods, local linear estimation and inverse probability weighting, as well as against the performance of a full cohort model assuming complete availability of the expensive covariate, via extensive simulation studies. Finally, we demonstrate the performance of our method on a large cohort study, looking at the association between oxidative stress and colorectal cancer.

4.3 Motivating Example

We were motivated by an application to data from a large cohort study, looking particularly at colorectal cancer incidence as it relates to antioxidant supplementation and oxidative stress. Conventional wisdom says that the antioxidant effects of supplements such as vitamin C, E, and selenium should lower risk of cancer by reducing cancer-causing oxidative stress, supported by early in vitro studies (Bjelakovic et al., 2004; Klein and Thompson Jr, 2011; Blot et al., 1993; Group, 1994; Gaziano et al., 2009; Omenn et al., 1996; Cole et al., 2007; Pais and Dumitraşcu, 2013) linking oxidative damage to carcinogenesis. However, large randomized controlled trials of antioxidants as cancer prevention agents have demonstrated no clear benefit and even a potential link to increased cancer risk. Recent in vitro evidence (Genestra, 2007; Chandel and Tuveson, 2014; Brunet et al., 2004; DeNicola et al., 2011; Kh and Ryan, 2009; Kops et al., 2002; Piskounova et al., 2015) has suggested that oxidative stress may shift from being pro- to anti-carcinogenic over the course of tumorigenesis as cancer cells develop adaptive mechanisms against oxidative damage.

In a pilot logistic regression analysis using this data, a bidirectional association was found between oxidative stress and CRC risk, demonstrating both beneficial and harmful effects that were time-dependent. More specifically, high oxidative stress was associated with an increased risk of CRC in early phases of cancer development, and a decreased risk in later phases. This pilot study also found that among patients with more advanced cancer, levels of oxidative stress were lower. This analysis was made possible because the study collected urine samples at baseline from all participants, and measured levels of oxidative stress present in these samples (via a biomarker) for those who developed colorectal cancer, along with sampled controls following a nested case control approach sampled 1-1. We build on this pilot study by fitting our proposed two-phase B-splines method in a Cox model that can incorporate Phase I information alongside the available Phase II biomarker data in Section 4.7, but begin with a complete description of our method and evaluation of its performance via simulation studies.

4.4 Methods

4.4.1 Efficient Two-Phase Estimation

We first describe our proposed two-phase estimation approach. Let T denote the event time, X the expensive covariate with a time-varying effect, and $\mathbf{Z}$ the vector of inexpensive covariates with time-fixed effects. We assume that the cumulative hazard function of the event time T conditional on covariates X and $\mathbf{Z}$ takes the form $\int_0^t \exp\left\{\beta(s)X+\gamma^{\mathrm{T}}\mathbf{Z}\right\}d\Lambda(s)$, where $\beta(\cdot)$ is the time-varying regression coefficient of X, γ is the time-fixed regression coefficient of $\mathbf{Z}$, and $\Lambda(\cdot)$ is an unspecified positive increasing function. In the presence of right censoring, we observe Y and Δ instead of T, where $Y=\min(T,C)$, $\Delta=I(T\leq C)$, C is the censoring time on T, and $I(\cdot)$ is the indicator function.

Let $P(\cdot|\cdot)$ denote a conditional density function. If $(Y,\Delta,X,\mathbf{Z})$ is observed for all n subjects in the study, then the inference on $\beta(\cdot)$, γ, and $\Lambda(\cdot)$ is typically based on the likelihood $\prod_{i=1}^n P(Y_i,\Delta_i|X_i,\mathbf{Z}_i)$. Under the two-phase design, however, only $(Y,\Delta,\mathbf{Z})$ is measured on all n subjects in Phase I, and X is measured for a sub-sample of size n_2 in Phase II. Let R be the selection indicator for the measurement of X in Phase II. We assume that the distribution of $(R_1,\ldots,R_n)$ depends on $(Y_i,\Delta_i,X_i,\mathbf{Z}_i)$ $(i=1,\ldots,n)$ only through the Phase I data $(Y_i,\Delta_i,\mathbf{Z}_i)$ $(i=1,\ldots,n)$. This assumption implies that the data on X are missing at random, such that the joint distribution of $(R_1,\ldots,R_n)$ conditional on $(Y_1,\Delta_1,\mathbf{Z}_1,\ldots,Y_n,\Delta_n,\mathbf{Z}_n)$ can be disregarded in the likelihood inference of $\beta(\cdot)$, γ, and $\Lambda(\cdot)$. Following Zeng and Lin (2014), we further assume that the censoring time C is independent of T given $(X,\mathbf{Z})$ among subjects with $R=1$ and independent of T and X given $\mathbf{Z}$ among subjects with $R=0$. In this situation, the observed-data likelihood takes the form

$$\sum_{i=1}^n R_i\left\{\log P(Y_i,\Delta_i|X_i,\mathbf{Z}_i)+\log P(X_i|\mathbf{Z}_i)\right\}+\sum_{i=1}^n(1-R_i)\log\left\{\int P(Y_i,\Delta_i|x,\mathbf{Z}_i)P(x|\mathbf{Z}_i)dx\right\}, \tag{4.1}$$

where

$$P(Y_i,\Delta_i|X_i,\mathbf{Z}_i)\propto\left[\Lambda'(Y_i)\exp\{\beta(Y_i)X_i+\gamma^{\mathrm{T}}\mathbf{Z}_i\}\right]^{\Delta}\exp\left[-\int_0^{Y_i}\exp\left\{\beta(t)X_i+\gamma^{\mathrm{T}}\mathbf{Z}_i\right\}d\Lambda(t)\right],$$

where $f'(x)=df(x)/dx$. Our main interest lies in the inference of $\beta(\cdot)$.

We use non-parametric maximum likelihood estimation and sieve approximation to maximize expression (4.1). Firstly, we approximate $\beta(\cdot)$ on $[0,\tau]$ using B-splines, where τ is the study duration; that is, we assume

$$\beta(t) = \sum_{l=1}^{d_n} \alpha_l A_l^w(t),$$

where $A_l^w(\cdot)$ is the lth B-spline basis function of order w on $[0, \tau]$, d_n is the total number of functions in this B-spline basis, and α_j is the coefficient for $A_l^w(\cdot)$ $(l = 1, \ldots, d_n)$ in the B-spline approximation of $\beta(\cdot)$. Secondly, we estimate $\Lambda(\cdot)$ by a step function with jumps only at the observed Y_i with $\Delta_i = 1$ $(i = 1, \ldots, n)$. Let λ_i be the jump size of $\Lambda(\cdot)$ at Y_i. We have $\lambda_i > 0$ and $= 0$ when $\Delta_i = 1$ and 0, respectively. Finally, if $\mathbf{Z}$ is discrete, then for each distinct observed $\mathbf{Z} = \mathbf{Z}$, we estimate $P(X|\mathbf{Z})$ by a discrete probability function on the distinct observed values of X, denoted by $x_1, \ldots, x_m$ $(m \leq n_2)$, where m is the total number of distinct values of X (i.e., m increases with n_2). If $\mathbf{Z}$ contains continuous components, then this nonparametric estimation procedure becomes infeasible because only a small number of observations on X are associated with each $\mathbf{Z}$. In this situation, we approximate $P(x|\mathbf{Z}_i)$ and $\log P(x|\mathbf{Z}_i)$ in expression (4.1) by

$$P(x|\mathbf{Z}_i) = \sum_{k=1}^{m} I(x = x_k) \sum_{j=1}^{s_n} B_j^q(\mathbf{Z}_i) p_{kj},$$

and

$$\sum_{k=1}^{m} I(x = x_k) \sum_{j=1}^{s_n} B_j^q(\mathbf{Z}_i) \log p_{kj},$$

respectively, where $B_j^q(\cdot)$ is the jth B-spline basis function of order q, s_n is the total number of functions in the B-spline basis, and p_{kj} is the coefficient of $B_j^q(\mathbf{Z}_i)$ at x_k $(k = 1, \ldots, m;\ j = 1, \ldots, s_n)$ in the B-spline approximation of $P(x|\mathbf{Z}_i)$. Details about the construction of the B-spline bases $\{A_l^w(\cdot)\}_{l=1}^{d_n}$ and $\{B_j^q(\cdot)\}_{j=1}^{s_n}$ and guidelines about the choices of (w, d_n) and (q, s_n) can be found in Chen et al. (2013) and Tao et al. (2017), respectively. In practice, w and q are typically chosen to be less than or equal to four, which corresponds to cubic splines, and d_n and s_n are determined by the Phase I sample size n. We note that p_{kj}

needs to satisfy the constraints

$$\sum_{k=1}^{m} p_{kj} = 1 \text{ and } p_{kj} \geq 0 \ (k = 1, \ldots, m;\ j = 1, \ldots, s_n) \tag{4.2}$$

because $P(x|\mathbf{Z}_i)$ is a conditional probability function. Consequently, the observed-data log-likelihood can be rewritten as

$$\begin{aligned} l_n(\theta, \{\lambda_i\}, \{p_{kj}\}) =& \sum_{i=1}^{n} R_i \left\{ \log P(Y_i, \Delta_i | X_i, \mathbf{Z}_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(X_i = x_k) B_j^q(\mathbf{Z}_i) \log p_{kj} \right\} \\ &+ \sum_{i=1}^{n} (1 - R_i) \log \left\{ \sum_{k=1}^{m} P(Y_i, \Delta_i | x_k, \mathbf{Z}_i) \sum_{j=1}^{s_n} B_j^q(\mathbf{Z}_i) p_{kj} \right\}, \end{aligned} \tag{4.3}$$

where

$$\begin{aligned} P(Y_i, \Delta_i | x, \mathbf{Z}_i) \propto & \left\{ \lambda_i \exp\left(\sum_{l=1}^{d_n} \alpha_l A_l^w(Y_i) x + \gamma^{\mathrm{T}} \mathbf{Z}_i \right) \right\}^{\Delta} \\ & \times \exp\left\{ - \sum_{i'=1}^{n} I(Y_{i'} \leq Y_i) \lambda_{i'} \exp\left(\sum_{l=1}^{d_n} \alpha_l A_l^w(Y_{i'}) x + \gamma^{\mathrm{T}} \mathbf{Z}_i \right) \right\}, \end{aligned}$$

We aim to maximize expression (4.3) under the constraints (4.2).

It is difficult to maximize expression (4.3) directly because of the intractable form of the second term. Following Tao et al. (2017), we solve this maximization problem by artificially creating a latent variable U for subjects with $R = 0$ such that U takes values on $1/s_n, 2/s_n, \ldots, 1$ and satisfies the equations

$$\begin{aligned} P(U = j/s_n | \mathbf{Z}) =& B_j^q(\mathbf{Z}), \\ P(X = x_k | \mathbf{Z}, U = j/s_n) =& P(X = x_k | U = j/s_n) = p_{kj}, \\ P(Y, \Delta | X, \mathbf{Z}, U) =& P(Y, \Delta | X, \mathbf{Z}). \end{aligned}$$

This step is essential because it enables us to interpret $\sum_{j=1}^{s_n} B_j^q(\mathbf{Z}) p_{kj}$ as $P(X = x_k | \mathbf{Z})$ for subjects with $R = 0$. Hence, the second term in expression (4.3) is equivalent to the log-likelihood of $(Y_i, \Delta_i, \mathbf{Z}_i)$, assuming that the complete data consist of $(Y_i, \Delta_i, X_i, \mathbf{Z}_i, U_i)$ but with X_i and U_i missing.

The maximization of expression (4.3) is carried out through an EM-algorithm, where (X,U) for subjects with $R=0$ are treated as missing data. The complete-data log-likelihood is

$$
\begin{aligned}
&\sum_{i=1}^{n} R_i \left\{ \log P(Y_i,\Delta_i|X_i,\mathbf{Z}_i) + \sum_{k=1}^{m}\sum_{j=1}^{s_n} I(X_i=x_k) B_j^q(\mathbf{Z}_i) \log p_{kj} \right\} \\
&+ \sum_{i=1}^{n} (1-R_i) \left\{ \log P(Y_i,\Delta_i|X_i,Z_i) + \log P(X_i|U_i) + \log p(U_i|\mathbf{Z}_i) \right\} \\
=& \sum_{i=1}^{n} R_i \left\{ \log P(Y_i,\Delta_i|X_i,\mathbf{Z}_i) + \sum_{k=1}^{m}\sum_{j=1}^{s_n} I(X_i=x_k) B_j^q(\mathbf{Z}_i) \log p_{kj} \right\} \\
&+ \sum_{i=1}^{n} (1-R_i) \left\{ \sum_{k=1}^{m} I(X_i=x_k) \log P(Y_i,\Delta_i|x_k,\mathbf{Z}_i) \right. \\
&\left. + \sum_{k=1}^{m}\sum_{j=1}^{s_n} I(X_i=x_k, U_i=j/s_n) \log p_{kj} + \sum_{j=1}^{s_n} I(U_i=j/s_n) \log B_j^q(\mathbf{Z}_i) \right\}.
\end{aligned}
$$

Let $\theta = (\alpha_1,\ldots,\alpha_{d_n},\gamma^{\mathrm{T}})^{\mathrm{T}}$. We start with the following initial values: $\widehat{\theta}^{(0)} = \mathbf{0}$, $\widehat{\lambda}_i^{(0)} = \Delta_i / \left(\sum_{i'=1}^{n} \Delta_i\right)$, and $\widehat{p}_{kj}^{(0)} = \left\{\sum_{i=1}^{n} R_i I(X_i=x_k) B_j^q(\mathbf{Z}_i)\right\} / \left\{\sum_{i=1}^{n} R_i B_j^q(\mathbf{Z}_i)\right\}$.

In the E-step of the $(t+1)$th iteration, we calculate the conditional expectations of $I(X_i=x_k, U_i=j/s_n)$ and $I(X_i=x_k)$ given $(Y_i,\Delta_i,\mathbf{Z}_i)$, $x_1,\ldots,x_m$, evaluated at $\widehat{\theta}^{(t)}$, $\widehat{\lambda}_1^{(t)},\ldots,\widehat{\lambda}_n^{(t)}$, $\widehat{p}_{11}^{(t)},\ldots,\widehat{p}_{ms_n}^{(t)}$, denoted as $\widehat{\psi}_{kji}^{(t+1)}$ and $\widehat{q}_{ik}^{(t+1)}$, respectively. That is,

$$
\widehat{\psi}_{kji}^{(t+1)} = \begin{cases} I(X_i=x_k) B_j^q(Z_i),\ R_i=1, \\ \dfrac{P(Y_i,\Delta_i|x_k,\mathbf{Z}_i) B_j^q(\mathbf{Z}_i) \widehat{p}_{kj}^{(t)}}{\sum_{k'=1}^{m} P(Y_i,\Delta_i|x_{k'},\mathbf{Z}_i) \sum_{j'=1}^{s_n} B_{j'}^q(\mathbf{Z}_i) \widehat{p}_{k'j'}^{(t)}},\ R_i=0, \end{cases}
$$

$$
\widehat{q}_{ik}^{(t+1)} = \begin{cases} I(X_i=x_k),\ R_i=1, \\ \dfrac{P(Y_i,\Delta_i|x_k,\mathbf{Z}_i) \sum_{j'=1}^{s_n} B_{j'}^q(\mathbf{Z}_i) \widehat{p}_{kj'}^{(t)}}{\sum_{k'=1}^{m} P(Y_i,\Delta_i|x_{k'},\mathbf{Z}_i) \sum_{j'=1}^{s_n} B_{j'}^q(\mathbf{Z}_i) \widehat{p}_{k'j'}^{(t)}},\ R_i=0. \end{cases}
$$

In the M-step of the $(t+1)$th iteration, we update $\widehat{\theta}^{(t+1)}$ and $\widehat{\lambda}_i^{(t+1)}$ $(i = 1, \ldots, n)$ by maximizing

$$\sum_{i=1}^{n}\sum_{k=1}^{m} \widehat{q}_{ik}^{(t+1)} \log P(Y_i, \Delta_i | x_k, \mathbf{Z}_i), \tag{4.4}$$

which is a weighted likelihood for the Cox model with time-varying effects. Let

$$\theta_{ik}(t) = (A_1^w(t)x_k, \ldots, A_{d_n}^w(t)x_k, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}},$$

and

$$G_{ik}(t) = \exp\left(\sum_{l=1}^{d_n} \alpha_l A_l^w(t)x_k + \gamma^{\mathrm{T}}\mathbf{Z}_i\right).$$

We update $\widehat{\theta}^{(t+1)}$ by solving

$$\mathbf{S}(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{m} \widehat{q}_{ik}^{(t+1)} \Delta_i \left\{ \theta_{ik}(Y_i) - \frac{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)\theta_{i'k}(Y_i)}{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)} \right\}$$

using the one-step Newton–Raphson algorithm. That is, we update $\widehat{\theta}^{(t+1)}$ by

$$\widehat{\theta}^{(t+1)} = \widehat{\theta}^{(t)} + \left\{\Omega\left(\widehat{\theta}^{(t)}\right)\right\}^{-1} S\left(\widehat{\theta}^{(t)}\right),$$

where

$$\Omega(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{m} \widehat{q}_{ik}^{(t+1)} \Delta_i \left[\frac{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)\theta_{i'k}(Y_i)^{\otimes 2}}{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)} \right. \\ \left. - \frac{\left\{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)\theta_{i'k}(Y_i)\right\}^{\otimes 2}}{\left\{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)\right\}^2} \right]$$

and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$. We update $\widehat{\lambda}_i^{(t+1)}$ by using the Breslow estimator such that

$$\widehat{\lambda}_i^{(t+1)} = \frac{\Delta_i}{\sum_{i'=1}^{n}\sum_{k=1}^{m} I(Y_{i'} \geq Y_i)\widehat{q}_{i'k}^{(t+1)} G_{i'k}(Y_i)}.$$

We observe that $\widehat{\lambda}_i^{(t+1)} > 0$ and $= 0$ when $\Delta_i = 1$ and 0, respectively. We update $\widehat{p}_{kj}^{(t+1)}$ $(k = 1, \ldots, m;$ $j = 1, \ldots, s_n)$ by maximizing

$$\sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{j=1}^{s_n} \left\{ R_i \widehat{q}_{ik}^{(t+1)} B_j^q(\mathbf{Z}_i) + (1 - R_i) \widehat{\psi}_{kji}^{(t+1)} \right\} \log p_{kj},$$

such that

$$\widehat{p}_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{ R_i \widehat{q}_{ik}^{(t+1)} B_j^q(\mathbf{Z}_i) + (1 - R_i) \widehat{\psi}_{kji}^{(t+1)} \right\}}{\sum_{k'=1}^{m} \sum_{i=1}^{n} \left\{ R_i \widehat{q}_{ik'}^{(t+1)} B_j^q(\mathbf{Z}_i) + (1 - R_i) \widehat{\psi}_{k'ji}^{(t+1)} \right\}}.$$

We observe that $\widehat{p}_{kj}^{(t+1)}$ satisfies the two constraints in expression (4.2).

We iterate between the E-step and M-step until

$$\left\| \widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)} \right\|_1 + \sum_{i=1}^{n} \left| \widehat{\lambda}^{(t+1)} - \widehat{\lambda}^{(t)} \right| + \sum_{k=1}^{m} \sum_{j=1}^{s_n} \left| \widehat{p}_{kj}^{(t+1)} - \widehat{p}_{kj}^{(t)} \right| < 10^{-4}$$

to obtain the sieve maximum likelihood estimator (SMLE) $\widehat{\theta}$, $\widehat{\lambda}_i$ $(i = 1, \ldots, n)$, and $\widehat{p}_{kj}$ $(k = 1, \ldots, m;$ $j = 1, \ldots, s_n)$.

To obtain the variance estimate of $\widehat{\theta}$, we use the profile likelihood method proposed by Murphy and van der Vaart (2000). By verifying the smoothness conditions of Theorem 1 in Murphy and van der Vaart (2000), it can be shown that the negative inverse of the Hessian matrix of the profile likelihood function $pl(\theta) = \max_{\{\{\lambda_i\}, \{p_{kj}\}\}} l_n(\theta, \{\lambda_i\}, \{p_{kj}\})$ is a consistent estimator for the limiting covariance matrix of $n^{1/2}(\widehat{\theta} - \theta)$. In practice, we obtain the value of $pl(\theta)$ by holding θ fixed in the EM algorithm and obtaining the value of $l_n(\theta, \{\lambda_i\}, \{p_{kj}\})$ at convergence. We estimate the covariance matrix of $\widehat{\theta}$ by the negative inverse of the matrix whose (k, l)th element is

$$h_n^{-2} \left\{ pl(\widehat{\theta} + e_k h_n + e_l h_n) - pl(\widehat{\theta} + e_k h_n) - pl(\widehat{\theta} + e_l h_n) + pl(\widehat{\theta}) \right\},$$

where e_k is the kth canonical vector, and h_n is a constant of the order $n^{-1/2}$.

4.4.2 Inverse Probability Weighting

We chose to compare our two-phase estimation method to inverse probability weighting (IPW), as this can be applied to case-cohort, nested case control, and mixed sampling designs (Støer and Samuelsen, 2016, 2012). With IPW, the Cox model is estimated using a weighted partial log-likelihood, where for a given subject i with weight w_i the weighted partial log-likelihood (Weyer and Binder, 2015) is defined as

$$l(\beta) = \sum_{i=1}^{n} \Delta_i w_i \left\{ \beta^T \mathbf{X}_i - \log\left(\sum_{q=1}^{n} I(Y_i \leq Y_q) w_q \exp(\beta \mathbf{X}_k) \right) \right\} .$$

The calculation of the weight w_i is dependent upon the sampling used by a given study. For our purposes we consider case-cohort sampling, nested case control sampling, and the special case where a study uses a combination of the two (referred to as "mixed" sampling), such as when combining multiple study sets that may have used different sampling methodologies. In the case-cohort setting, sampling weights are decided by the sampling proportion used. For a given sampling probability π, weights for sampled controls will be $w_i = \frac{1}{\pi}$, and weights for cases will be $w_i = 1$ (Støer and Samuelsen, 2012). In the nested case control setting, Kaplan-Meier type weights, developed by Samuelsen (1997), are used, where the sampling probability is first calculated as

$$\pi_i = 1 - \prod_{0<t_j<t_i} \left\{ 1 - \frac{M}{n(t_j) - 1} \right\}$$

where $n(t_j)$ is the number at risk in the cohort at time j, M is the number of sampled controls per case, and t_i is the event (or censoring time) of subject i. Again, all cases have a sampling weight of 1, and controls will have a sampling weight $w_i = \frac{1}{\pi_i}$. In the mixed case, the total study population is first divided into subsets based on the study of origin, and sampling weights are then calculated within sub-study based on sampling design. Once this is done, the subsets are recombined to re-form the total study population.

Once sampling weights are calculated, the weighted partial log-likelihood is used to estimate $\beta(t)$ and γ on only those with complete observed data, or where $R = 1$. The robust variance estimator is used to account for the weighted estimation of the partial log-likelihood.

4.4.3 Local Linear Estimation for Nested Case Control Data

As an additional comparison method specific to nested case control studies, we considered local linear estimation following the procedure outlined in Liu et al. (2010). Briefly, in this approach $\beta(t)$ is estimated locally at pre-specified time points by calculating the partial likelihood in a neighborhood surrounding each time point t. More specifically, for a specific time t, let $R(t) = \{i : T_i \geq t\}$ denote the risk set. With nested case control sampling, each case has M controls sampled without replacement from the risk set at that case's failure time, exclusive of the case itself. So, for a case i, R_i^* denotes the indices of the M selected controls, and thus we can define $R_i = R_i^* \cup \{i\}$, the union of the control and case indices. We denote the individual coefficient estimates as $\mathbf{a}(s)$, where s is in a neighborhood of t, and allow all covariates in the model to have time-varying effects during estimation, such that $(\beta(t), \gamma) = \mathbf{a}(t) = (a_1(t), a_2(t), \ldots, a_p(t))$, where γ was of length $p-1$, p the total number of covariates including X. Now, assuming that the coefficient function $\mathbf{a}(s)$ has a continuous second derivative within the neighborhood of t and that $\mathbf{X}(t)$ (where $\mathbf{X} = (X, \mathbf{Z})$) is a locally bounded predictable process, then for s in a neighborhood of t and $v = 1, 2, \ldots, p$, by Taylor's expansion we can approximate $a_v(s)$ by

$$a_v(s) \approx a_v(t) + a_v'(t)(s-t)$$

Define $\beta = (a_1(t), \ldots, a_p(t), a_1'(t), \ldots, a_p'(t))^T$ and $\tilde{\mathbf{X}}_i(u, u-t) = \mathbf{X}_i(u) \otimes (1, u-t)^T$ with $\otimes$ being the Kronecker product. Further, let $h = h_n > 0$ be a bandwidth parameter controlling the size of a local neighborhood and $K(\cdot)$ be a kernel function that smoothly down-weights the contribution of remote data points. Then, by incorporating localized weights and risk set indicators, the local linear partial likelihood function is obtained by

$$l(\beta) = \sum_{q=1}^{n} \int_0^{\tau} K_h(u-t) \left[\tilde{\mathbf{X}}_q(u, u-t)^T \beta - \log \left\{ \sum_{j \in R_q}^{M+1} \exp(\tilde{\mathbf{X}}_j(u, u-t)^T \beta) \right\} \right] dN_q(u) \tag{4.5}$$

where $K_h(\cdot) = K(\cdot/h)/h$. Let $\hat{\beta}$ be the maximum likelihood estimate of (4.5) with respect to β. Then the local linear partial maximum likelihood estimate of $\mathbf{a}(t)$, $\hat{\mathbf{a}}(t)$, is the first p components of $\hat{\beta}$, while the last p components estimate the derivative of $\mathbf{a}(t)$. The very first component of this vector, $\hat{a}_1(t)$, is of interest for this manuscript, as our goal is estimation of the time-varying coefficient on the expensive covariate X in particular. For further details, including details on the variance of $\hat{\mathbf{a}}(t)$, can be found in Liu et al. (2010).

4.5 Simulation Study

4.5.1 Data Simulation

We evaluate the performance of our method via simulation under several data conditions. For each condition, 2,000 subjects were simulated with a hazard funciton for subject $i \in (1, 2, \ldots, 2000)$ defined as

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(\beta(t)X_i + \gamma Z_i)$$

where X_i and Z_i were either both binary or both continuous covariates, and X and Z are correlated with one another. In addition to different types of X and Z, we also varied the target event rate (moderate or low) and level of correlation between X and Z (moderate or high). The resultant values for $\lambda_0(t)$, Z, and X can be found in Table 4.1. Across all simulations, $\gamma = 0.15$. Censoring times were generated from the exponential distribution with hazard function $\exp(0.1 + 0.1Z_i)$, truncated by 0.05 and the largest follow-up time of $\tau = 1.8$.

Table 4.1: Specifications for data simulations to produce desired event rates and correlation between X and Z.

Data type	Event rate	Correlation	$\lambda_0(t)$	Z	X
Binary	$12-15\%$	$0.3 < \rho < 0.4$	0.20	$Z \sim \text{Bin}(0.5)$	$X\|Z = 0 \sim \text{Bin}(0.3)$ $X\|Z = 1 \sim \text{Bin}(0.7)$
Binary	$12-15\%$	$0.7 < \rho < 0.8$	0.20	$Z \sim \text{Bin}(0.5)$	$X\|Z = 0 \sim \text{Bin}(0.15)$ $X\|Z = 1 \sim \text{Bin}(0.85)$
Binary	$4-5\%$	$0.7 < \rho < 0.8$	0.07	$Z \sim \text{Bin}(0.5)$	$X\|Z = 0 \sim \text{Bin}(0.15)$ $X\|Z = 1 \sim \text{Bin}(0.85)$
Continuous	$12-15\%$	$0.3 < \rho < 0.4$	0.20	$Z \sim \text{Unif}(0,1)$	$X \sim 0.4Z + \text{Unif}(0,1)$
Continuous	$12-15\%$	$0.7 < \rho < 0.8$	0.20	$Z \sim \text{Unif}(0,1)$	$X \sim Z + \text{Unif}(0,1)$
Continuous	$4-5\%$	$0.7 < \rho < 0.8$	0.07	$Z \sim \text{Unif}(0,1)$	$X \sim Z + \text{Unif}(0,1)$

To evaluate performance of these methods across different shapes of $\beta(t)$, we considered five functional forms of $\beta(t)$ of differing complexity:

(a) $\beta(t) = 0.1$, a time-independent effect

(b) $\beta(t) = 0.1\log(t+1)$, a monotone increasing effect

(c) $\beta(t) = \log(0.7 + 2t\exp(-t^2))$, a non-monotonic effect with the highest hazard at the midpoint that changes direction gradually

(d) $\beta(t) = \log(1.2 - 0.9\sin(2t))$, a non-monotonic effect with the lowest hazard at the midpoint that changes direction sharply

(e) $\beta(t) = 0.3 - 0.5 \cdot (\sin(2t) + \cos(2\sqrt{3}\,t))$, a non-monotonic effect with changing curvature at both the midpoint and tails

The shapes of these coefficients across $t \in (0, 1.8)$ can be found in Figure 3.1. For each data set, subjects were chosen for inclusion in Phase II using one of three methods: case-cohort sampling, nested case control sampling, or mixed sampling. In the case-cohort setting, a random sample of 20% of the total population was taken, and then all additional cases not included in this random sample were added. In the nested case control setting, we chose $M = 1$ controls for each case. In the mixed setting, subjects were first randomly sampled to either case-cohort or nested case control status, with half of subjects going to each. From there, we repeated the above sampling approaches for each sub-population, and then re-combined the subjects to form the complete cohort of 2,000 for each simulation. Subjects who were sampled for Phase II have $R = 1$, and those who were not have $R = 0$. For subjects not sampled for Phase II, we set the value of X as missing.

4.5.2 Model Fitting

For both the two-phase and IPW fitting procedures, the B-spline term on X was fit with order 4 (degree 3) and 5 interior knots, placed at quantiles $(0.05, 0.275, 0.50, 0.725, 0.95)$ (Harrell et al., 2001). These settings were based on the "optimal" fits displayed in Chapter 3, Section 3.5.1. For the two-phase method, the B-spline on Z was fit with order 4 and 4 interior knots, placed at quantiles $(0.05, 0.35, 0.65, 0.95)$. For local linear estimation in the nested case control setting, a bandwidth of $h = 0.3$ was used, in alignment with the optimal as proposed by Liu et al. (2010) based on a sample size of $n = 2000$ as well as the results of our simulations in Chapter 3.

Models using B-splines were fit using the **coxph** function in the **survival** package version 3.1-8 in R version 3.6.0 (Therneau, 2015; R Core Team, 2020). To fit $\beta(t)$, the **tt** option was used, with X multiplied by the corresponding basis matrix over event times t. All B-spline basis matrices were generated using the **bs** function from the **splines** package version 3.6.3. In the IPW case, weights for case-cohort populations were calculated as $1/p$, where p is the sampling proportion (in our case 0.2). For nested case-control populations, weights were calculated as the inverse of the sampling probabilities from the **KMprob** function in the **multipleNCC** package version 1.2-2 (Støer and Samuelsen, 2016), following the approach described in Section 4.4.2. For all estimates of $\beta(t)$ found via local linear estimation, the local linear partial likelihood function was maximized using the **optim** function from the **stats** package version 4.0.5, setting the maximum iteration limit to 10,000. Optimization was performed via the Nelder-Mead method (Nelder and Mead, 1965). Both the IPW and local linear estimation models were fit to complete case data only.

As a gold standard comparison model, we also fit a model to the complete data on all 2,000 subjects, with X and Z available for all, following the structure outlined in Section 3.3.2.2 and fit with the same 5 knots and degree 3 as both the two-phase and IPW models. These models were fit to all relevant simulation settings (data type, event rate, correlation level, curve shape), and used to evaluate the relative efficiency of the two-phase method in estimating $\beta(t)$ across time.

The variance estimate of $\beta(t)$ for the two-phase fitting procedure was obtained by first using the profile likelihood method described in Section 4.4.1 to estimate the variance-covariance matrix of $\widehat{\theta}$, Σ_θ. Then, the elements corresponding the basis coefficients α, denoted Σ_α, were used along with the corresponding basis functions $\mathbf{A}(t)$, to get $\widehat{\text{Var}}(\beta(t)) = \mathbf{A}(t)\Sigma_\alpha\mathbf{A}(t)^T$. For IPW fits, robust variance estimates were calculated during fitting, and the resulting variance-covariance matrix was used alongside the corresponding basis functions to calculate $\widehat{\text{Var}}(\beta(t))$ as described. For the full cohort fits, the variance estimate from **coxph** (non-robust) was used to calculate $\widehat{\text{Var}}(\beta(t))$. Finally, for local linear estimation, the variance of $\beta(t)$ is calculated at each time point t following Liu et al. (2010), based on the second derivative of the local linear partial likelihood function.

4.5.3 Methods for Evaluating Simulation Results

We compared the two-phase fitting procedure against both IPW and local linear estimation in several ways. First, for each repetition j of each procedure under each simulation setting (event rate, correlation level, curve shape, and data type), where $j = 1, 2, \ldots, 400$, the estimate of $\beta(t)$ was evaluated for $t \in (0, 1.8)$ using the original knot locations and settings for the spline terms $A_l^w(t)$, in intervals of 0.01. For local linear estimation, estimates of $\beta(t)$ were calculated at intervals of 0.05 during fitting. Then, the median across simulations at each time point was calculated, so that for a given t we report $\tilde{\beta}(t)$ as the median of $\beta(t)$, a vector of length 400 where the jth element is $\hat{\beta}(t)_j$, the estimate from the jth simulation setting. Additionally, 95% confidence interval bounds for each $\hat{\beta}(t)$ were calculated as $\hat{\beta}(t)_j \pm 1.96 \cdot SE(\hat{\beta}(t)_j)$, based on the fitted variance of $\beta(t)$ described in Section 4.5.2. We reported the medians of the lower and upper bounds of these confidence intervals, which are reported for each value of t as the "Fitted SE" confidence intervals in our summary figures. We also calculate 95% empirical confidence intervals around $\tilde{\beta}(t)$ based on the empirical standard deviation of all $\hat{\beta}(t)$ estimates for each time point t, which we calculate as $\tilde{\beta}(t) \pm 1.96 \cdot SD(\beta(t))$. These are plotted in our summary figures as "Empirical SD" intervals. Plotting both allows us to visually evaluate the ability of the variance estimator in each method to capture the empirical variance in $\hat{\beta}(t)$.

Additionally, we compare methods based on coverage across time, where coverage at time point t is calculated as the percentage of simulation repetitions (out of 400) whose lower and upper "Fitted SE" bounds contain the true value of $\beta(t)$. We also compare based on mean square error, calculated per time point as $MSE(t) = \frac{1}{n_s}\sum_{j=1}^{n_s}(\beta(t) - \hat{\beta(t)}_j)^2$, where $n_s = 400$.

To compare the relative performance of our two-phase fitting procedure and IPW in particular, as these are both calculated for all sampling types as opposed to local linear estimation which is restricted to nested case control sampling, we look at the median relative efficiency of the estimates of $\beta(t)$ across time, calculated as $Var(\beta(t)_{ipw})/Var(\beta(t)_{twophase})$ and summarized as the median across the 400 repetitions for each setting, where values greater than 1 signify that the two-phase procedure is more efficient than IPW, and vice versa. We repeat this process for estimates of γ, the time-independent inexpensive covariate effect. Finally, the two-phase fits were compared against the full cohort models fit to the same simulated data sets in terms of the median relative efficiency of $\beta(t)$, calculated as $Var(\beta(t)_{full})/Var(\beta(t)_{twophase})$ across t

and summarized across the 400 repetition simulations for each setting.

All models were fit using the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) cluster on Intel Sandy Bridge architecture processors. Both IPW and local linear estimation fits were allotted 8 GB of system memory, while the two-phase fits were allotted 32 GB to account for the increased total data size during fitting as a result of incorporating all Phase I data. Computation time per fit j was recorded and compared across fit types and data settings as a measure of computational efficiency.

4.6 Simulation Results

All simulation results can be found in Figures 4.1 to 4.10. Each figure represents the combination of a data scenario (correlation and event rate, as described in Table 4.1) and a sampling type (among case-cohort, mixed, and NCC). The details for each can be found in the figure captions. Among Figures 4.1 to 4.9, all panels are arranged by data type on the rows (binary versus continuous) and curve shape across the columns (shapes (a) through (e)). The first panel of each figure, **A**, displays the estimated $\beta(t)$ curves from each method, along with both Fitted SE and Empirical SD confidence intervals. The second panel, **B**, displays both coverage probabilities on the left y-axis, plotted for all fits, as well as the relative efficiency of the B-spline fit to the IPW fit on the right y-axis. The dotted reference line displays the nominal 95% coverage level at 0.95 for the left y-axis and the equal efficiency level at 1.0 for the right y-axis. The final panel, **C**, demonstrates mean square error, plotted by fit type.

Figures 4.1 to 4.3 present the results of our simulations fit to a event rate of $12-15\%$ with a moderate level of correlation between X and Z of $0.3 < \rho < 0.4$. Under this data scenario, the differences between our two-phase B-splines fit (in purple) and the IPW fit (in orange) are slight, but in general the two-phase fit demonstrates slightly narrower confidence intervals and the corresponding slightly higher efficiency (with the relative efficiency values plotted in pink in panel **B** slightly above 1 for the majority of the observed time). The relative efficiency of the two-phase fit to the IPW fit dips below 1 at the latest observed times, but this occurs during the same interval of time over which the IPW confidence intervals drop below nominal coverage. Thus, although the variance of $\beta(t)$ in this time is higher for the two-phase fits, the IPW fits with lower variance also suffer from low coverage. Differences in MSE are slight, but tend to be in favor of the two-phase fits particularly in the mid-range of time. The local linear estimation fits in Figure

4.3 demonstrate wide confidence intervals and high mean square error across time, but particularly at the higher end of times t. Of note in this and the next data settings is the difference in confidence intervals between the Empirical SD and Fitted SE in the binary data case. This finding is consistent between the IPW and two-phase B-splines fits, and particularly apparent in the local linear estimation fits, and is likely related to difficult in estimating $\beta(t)$ in time ranges where few subjects with $X = 1$ experienced events.

Figures 4.4 to 4.6 pertain to our fits with an event rate of $12-15\%$ and a high level of correlation between X and Z of $0.7 < \rho < 0.8$. The relative efficiency of the two-phase B-splines approach to IPW (panel **B**) under this setting was higher than with a moderate level of correlation, with values consistently near 1.3 - 1.4 except for at the extreme ends of the time scale. Once again, the region of lower efficiency aligns with low coverage for the IPW confidence intervals. Both methods do well in approximating the true $\beta(t)$ curve in panel **A**, but in alignment with the increased efficiency of the two-phase estimator we see narrower confidence intervals. Finally, mean square error values (panel **C**) are comparable between the two methods, though the two-phase approach has slightly lower mean square error particularly in the continuous data setting. Again in Figure 4.6 we find that local linear estimation suffers from high mean square error and wide confidence intervals, and a large disparity between its empirical SD and fitted SE confidence intervals.

For the final grouping, Figures 4.7 to 4.9 summarize our fits in the rare disease case, with an event rate of $4-5\%$ and a high correlation between X and Z of $0.7 < \rho < 0.8$. Efficiency gains here (panel **B**) are comparable to the non-rare high correlation setting, with the same low efficiency region corresponding to low coverage for IPW in the right tail of the time range. Differences in mean square error (panel **C**) are particularly pronounced in the nested case control setting in Figure 4.9 for continuous X and Z. Note that for the rare disease case, the local linear estimation model was unable to fit to the reduced event rate due to issues inverting the Hessian matrix; as such, we have excluded the local linear estimation model from this figure.

Additionally, we present the relative efficiency in the estimation of γ between the two-phase B-splines and IPW fits. As before, rows reflect data type (binary or continuous) and columns curve shape. In each panel of the figure, sampling type is plotted from left to right as case-cohort (CC), mixed, and nested case control (NCC). The different colors of the points correspond to data scenarios, with green representing high correlation and an event rate of $12-15\%$, pink a moderate correlation and event rate of $12-15\%$, and blue a high correlation and event rate of $4-5\%$. Relative efficiency is consistently above 1, indicating that the two-phase estimation method is more efficient. For the rare disease setting in particular (blue), the nested case control approach demonstrates the largest gain in efficiency across all data types and curve shapes. When correlation between X and Z is only moderate and the event rate is $12-15\%$ (pink), sampling type does not seem to make a difference in terms of efficiency gains, though all relative efficiency values in this case are greater than at least 3.

Our comparison of efficiency in estimating $\beta(t)$ between our two-phase B-splines fit and the corresponding model fit to full data (no sampling) is in Figure 4.11. For both scenarios with high correlation between X and Z (panels **B** and **C**), we see comparable efficiency between the two-phase B-splines fit and the full cohort fit, with values consistently near 1. In the extreme tails, particularly for large values of t, the two-phase fit is more efficient than the full cohort fit, but this is likely due to the extreme tail behavior demonstrated in Chapter 3 when estimating time-varying effects using B-splines, and more of an empirical finding than a theoretical one. When the correlation between X and Z is moderate (panel **A**), for binary data the two methods are comparably efficient with values consistently around 1. For continuous data, the full cohort fit is more efficient than the two-phase fit, with median relative efficiency values averaging around 0.62 across time. Finally, we can compare computational efifciency for the two-phase, IPW, and local linear estimation fits. The IPW fits were the fastest across sampling types and data scenarios, averaging 2-3 seconds in binary case and 5-7 seconds in the continuous setting. The local linear estimation fits took 4-5 seconds in the binary case and 6-7 seconds in the continuous case. The two-phase fits took 2-3 minutes in the binary case and 1-2 hours in the most complex continuous case (curve shape (e)). This computation time difference is due in part to the underlying process behind the **tt** function used by **coxph** to fit time-varying coefficients, which involves the creating of a counting process dataset. While the IPW fit also used this function, the original datasets entering the model differ in number of rows, as the two-phase fit includes all Phase I participants and expands their records to a number of rows dictated by the number of unique values of X, while the IPW fits only include Phase II participants.

4.7 Real-World Application: Oxidative Stress and Cancer Risk

4.7.1 Data Source and Methods

As previously explained, we chose to illustrate the performance of our method on real-world data using data from a cancer cohort study to study the time-varying effect of antioxidant stress on cancer risk. We evaluate this hypothesis by fitting our two-phase B-splines method to the two-phase nested case control data available from this study, along with fitting the IPW method described previously with Kaplan-Meier type weights for nested case control studies as a comparator.

The full cohort, comprising Phases I and II, was 64,410 subjects. Among these, there were 684 cases (an event rate of 1%), and 718 total controls selected. These two groups combined formed the 1,402 subjects used in the IPW analysis. For the two-phase B-splines analysis, we were unable to fit to the entire population, as the required space for creation of the counting process matrix exceeded our maximum available RAM of 1 terabyte. Instead, we fit to two sub-sampled populations. In the first, we sub-sampled in a 3:1 ratio with the existing controls in Phase II from those only in Phase I, adding 2,154 subjects from Phase I to those in Phase II for a total sample size of 3,556. In the second, we sub-sampled in a 10:1 ratio, adding 7,180 subjects from Phase I to those in Phase II for a total sample size of 8,582. We fit the two-phase method to both of these sub-samples, and compare the performance of the two to see the trade-off of computational efficiency relative to the resultant estimate of $\beta(t)$.

Available study covariates included variables related to oxidative stress such as smoking status, BMI, Charleson comorbidity index, menopause status, intake of vitamins C, E, and selenium, and the routine use of antioxidant vitamins. Additionally, there were covariates related to cancer risk, including having a family history of colorectal cancer as well as routine use of NSAIDs, and general demographic variables on education level. Finally, there were variables relating to urine sample collection, including sample time, the proximity of the urine sample collection to the last meal consumed, and the proximity of the sample to the most recent consumption of antibiotics, vitamin supplements, and NSAIDs. The outcome of interest was time to cancer diagnosis, calculated as the difference between age at baseline and age at diagnosis. For controls, censoring time C was determined as the difference between age of enrollment and age at last follow-up. The covariate with a suspected time-varying effect was the value of the oxidative stress biomarker measurement.

As there were several candidate inexpensive covariates Z to use in the two-phase B-splines approach, we chose to use principal components (PCA) (Salem and Hussein, 2019) to reduce dimension prior to estimating the conditional probability function. To determine which covariates should be included in the PCA, we use Phase II data to fit a linear regression model with an outcome of oxidative stress and included all previously listed covariates as predictors. Any covariate with $p < 0.3$ was included in the PCA. The first component of the resulting PCA was used as $\mathbf{Z}$ in $P(x|\mathbf{Z})$, and a B-spline was created for this variable with degree 3 and 4 knots placed at quantiles $(0.05, 0.35, 0.65, 0.95)$, boundary knots over the range of the component, and with an intercept. From there the model fit proceeded as described in Sections 4.4.1 and 4.5.2, with the B-spline basis function used to approximate $\beta(t)$ fit with 5 knots (with the same knot placement as in Section 4.5.2), degree 3, and an intercept term. The value of the biomarker was rounded to two decimal places. All candidate covariates were included in the final model, including those not included in the PCA.

The IPW model for comparison was fit to the same set of covariates as the two-phase model, and with the same specifications for the B-spline basis used to approximate $\beta(t)$. Sampling weights were calculated with respect to the total population size of 64,410. Models were summarized by their estimate of $\beta(t)$ and the fitted standard error 95% confidence intervals surrounding them, and compared across the range of unique event times, from $t = 0.1$ to $t = 14.4$ years.

4.7.2 Results

The results of our fit can be found in Figure 4.12. In orange the IPW fit is plotted, in purple the two-phase B-splines fit with a 3:1 sub-sampling ratio to the controls, and in pink the fit with 10:1 sub-sampling. We can see that the estimated curve for $\beta(t)$ is consistent across all three methods, hovering at or slightly above 0 for the first 4 years before dipping below 0 between 4.5 and 10 years, and ultimately returning to 0 for the rest of the time scale. The difference in fits lies in the width of the confidence intervals. For the IPW fit, the 95% confidence interval contains 0 throughout the entire time scale, failing to capture any significant difference between 0 and $\beta(t)$. For both the 3:1 and 10:1 B-splines fits, however, the upper bound of the confidence interval dips below 0 between years 6 and 9, demonstrating a significant association between the value of the oxidative stress biomarker and time to diagnosis. Also importantly, the direction of this finding is in line with the pilot study, which found that oxidative stress shifted from pro- to anti-carcinogenic over time, indicated by $\beta(t)$ shifting from slightly positive to negative. While the two-phase B-splines fit had the increased inferential efficiency to detect a significant difference in this case, it did so at a trade-off in computational efficiency. The IPW fit, requiring 8 GB of RAM, took roughly 5 seconds to fit, while the two-phase B-splines fit with 3:1 sub-sampling on 32 GB took approximately 1 hour, and the fit with 10:1 sub-sampling on 64 GB took 18 hours. However, there was no significant difference in performance between these two sub-sampled fits, so it is possible that at 3:1 sub-sampling (which creates a pseudo 4:1 NCC sample) a threshold of efficiency is reached, and increasing sub-sampling is not necessary. Thus, the trade-off becomes about one hour of computing time in return for an increase in inferential efficiency leading to a significant result.

4.8 Discussion

We demonstrated via simulation the performance of our proposed two-phase B-splines approach to estimating time-varying effects in the Cox model under two-phase study designs. Over different data scenarios, sampling methods, and functional forms of $\beta(t)$ we found that our approach was able to capture the shape of $\beta(t)$ across time while achieving the nominal coverage probability of 0.95. This finding is consistent with Chapter 3, where we found that B-splines were able to maintain nominal coverage regardless of simulation setting. When compared to an IPW fit on the same simulated data, we found that our method had higher efficiency for the majority of the time under consideration. This gain in efficiency can be attributed to the fact that our approach incorporated all Phase I information as well as the Phase II information from those sampled. In doing so, it was between 3 and 6 times as efficient in estimating γ, for which it had complete information, and $10-40\%$ more efficient in estimating $\beta(t)$ across data settings and sampling schemes. Efficiency gains were particularly notable when X and Z were highly correlated, and when the disease of interest was rare (event rate $4-5\%$). In these cases, the conditional density estimation of X on Z provides more information than when there is less correlation between the two covariates.

Previous work using the SMLE for model estimation (Tao et al., 2017) demonstrated that it produced asymptotically efficient estimators, and our comparison to the full cohort model shown in Figure 4.11 aligns with this finding, with our two-phase fit near equivalent in efficiency to the full-cohort fit across time.

The lower efficiency of IPW is also consistent with the findings of Støer and Samuelsen (2012), who noted that the IPW method lacked in efficiency for the nested case control setting when compared to methods evaluating a complete likelihood that included Phase I information from all subjects, such as that described in Saarela et al. (2008). However, their method required parametric specification of both the baseline hazard and the conditional distribution of X given Z, while our method requires no such parametric assumptions. Additionally, both Støer and Samuelsen (2012) and Saarela et al. (2008) were focused on the time-independent effect setting, while our work extends to time-varying coefficient effects.

Our comparison to local linear estimation for NCC studies found that, when compared to our two-phase approach as well as to IPW, local linear estimation demonstrated high variability, particularly when comparing the empirical confidence intervals to those calculated using the fitted standard errors. Additionally, local linear estimation was unable to fit to our rare disease data due to issues inverting the Hessian matrix. As Liu et al. (2010) did not consider the rare event setting (focusing instead on event rates of $10-13\%$, it is hard to say if this finding is consistent with theirs. A wider bandwidth setting may reduce the error in these fits, however as demonstrated in Chapter 3 this would likely come at the detriment of coverage probability, particularly at points of articulation in the shape of $\beta(t)$.

Of note is that our method is applicable to multiple types of sampling procedures, with the only assumption of the current work being that sampling is dependent on only one outcome. As explained in Tao et al. (2017), if sampling were dependent on multiple outcomes then these should all be considered simultaneously in a multivariate regression model to obtain valid inference. Our method could be extended to multiple outcomes survival analysis by replacing the conditional density $P(Y_i,\Delta_i|X_i,\mathbf{Z}_i)$ with $P(\mathbf{Y}_i,\Delta_i|X_i,\mathbf{Z}_i)$, where $\mathbf{Y}_i$ is the multivariate outcome vector given covariates X_i and $\mathbf{Z}_i$.

While our simulations only dealt with a single Z covariate, as demonstrated in our application in Section 4.7 our method can incorporate multiple inexpensive covariates, provided dimension reduction is possible. This is particularly necessary when Z contains multiple continuous components. With just two continuous components, a multivariate B-spline on $\mathbf{Z}$ may be possible for the conditional density function, as explained in Tao et al. (2017); however, beyond two continuous components, dimension reduction via PCA or a similar method is necessary to make computation feasible. In addition to dimension reduction, the computational intensity of our two-phase method can be reduced as needed by rounding event times to reduce the total number of unique times and rounding X (if continuous) to reduce the unique values across which p_{kj} needs to be estimated. Additionally, with particularly large complete datasets, such as that used in our example on CRC risk, sub-sampling of the Phase I controls can be used to reduce computational demand while still producing efficient results, with a 3:1 sub-sampling ratio producing comparable results to a 10:1 ratio.

Acknowledgments

This work was conducted in part using the resources of the ACCRE cluster at Vanderbilt University, Nashville, TN.

Funding

- E.A.S.W. was supported by NIH grant U24CA194215.
- R.T. was supported by NIH grant R01AI131771 and the VUMC Biostatistics Faculty Development Award.
- Q.C. was supported by NIH grants U24CA194215.

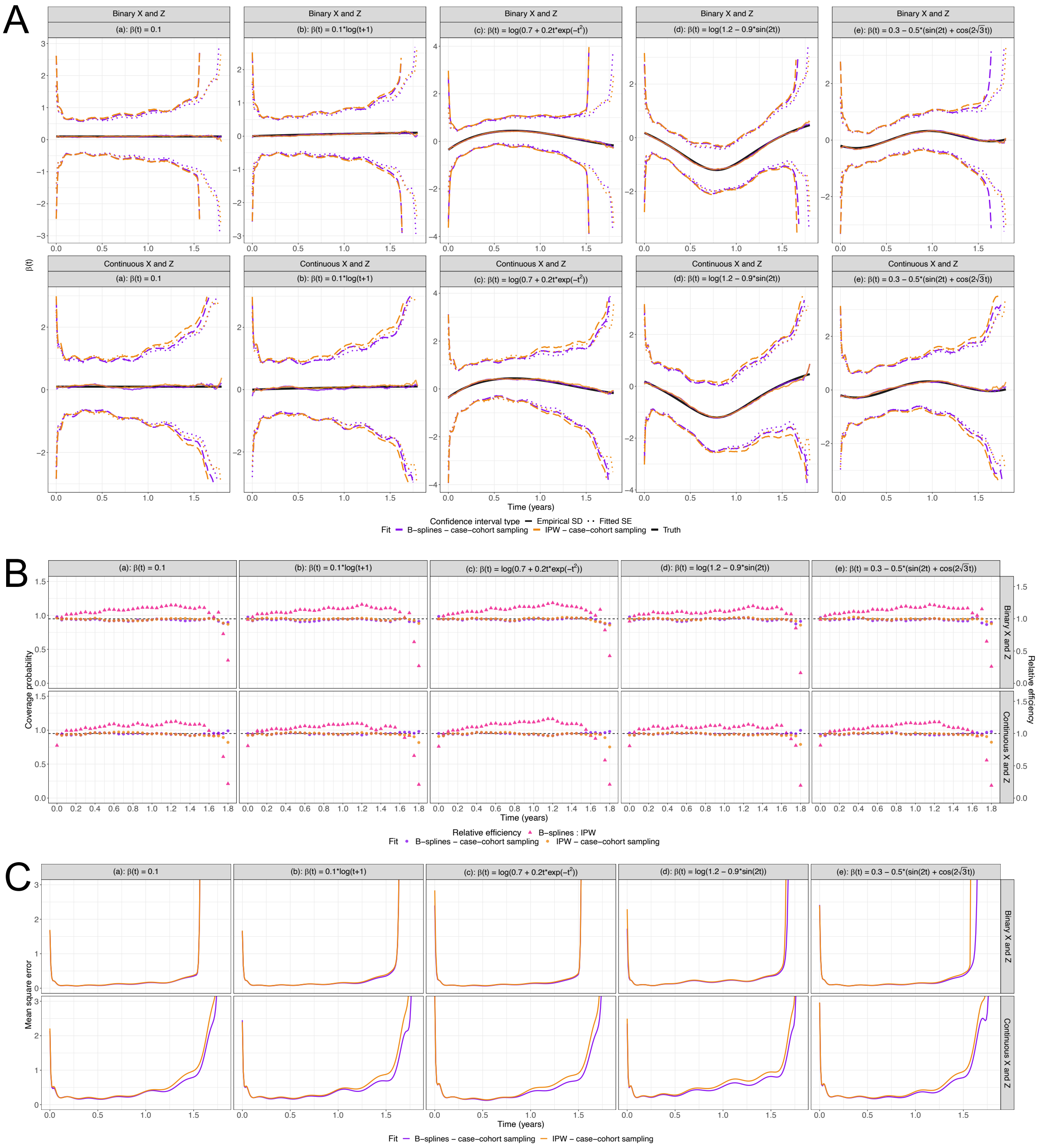

Figure 4.1: Simulation results for moderate correlation, case-cohort sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

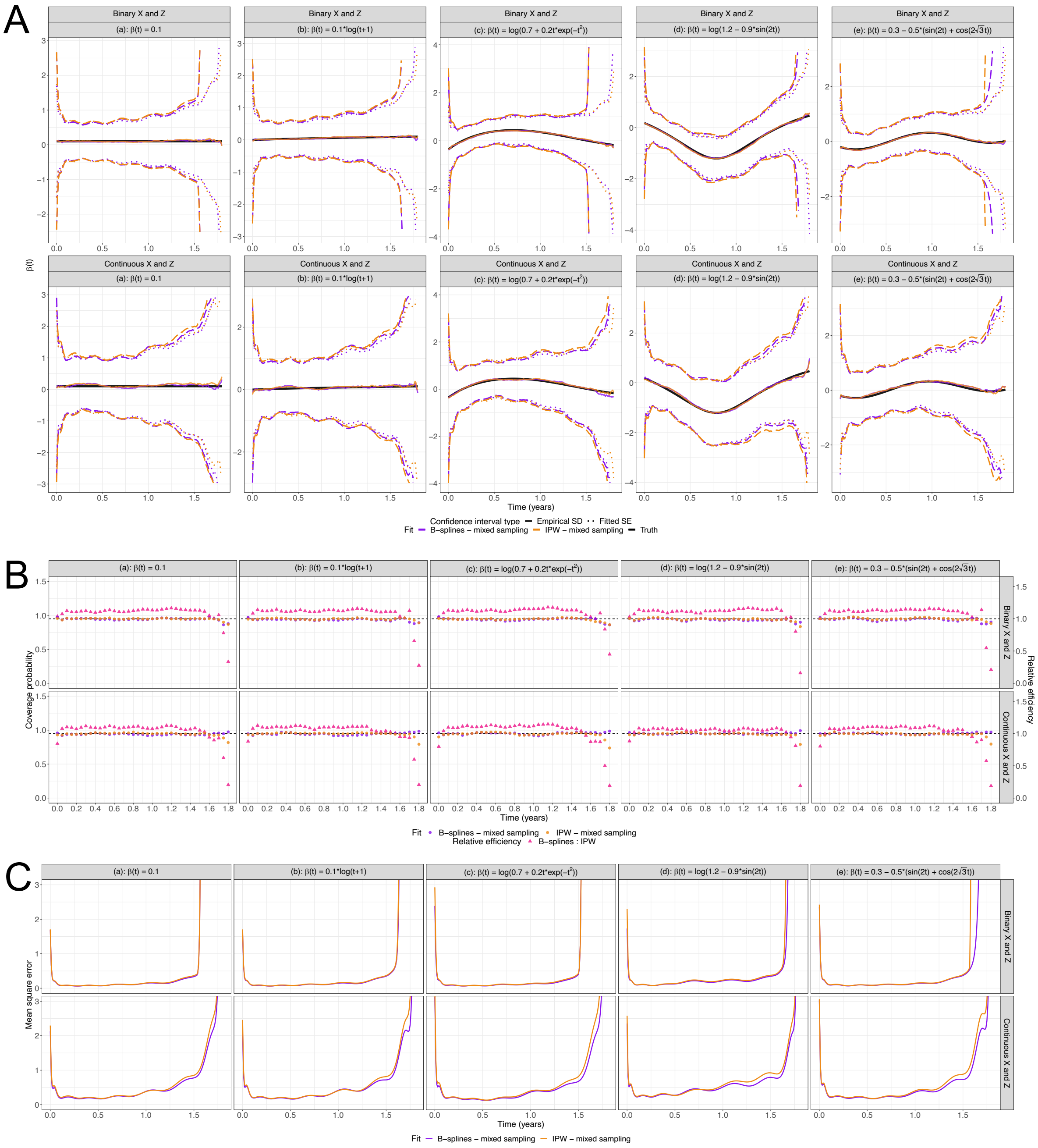

Figure 4.2: Simulation results for moderate correlation, mixed sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

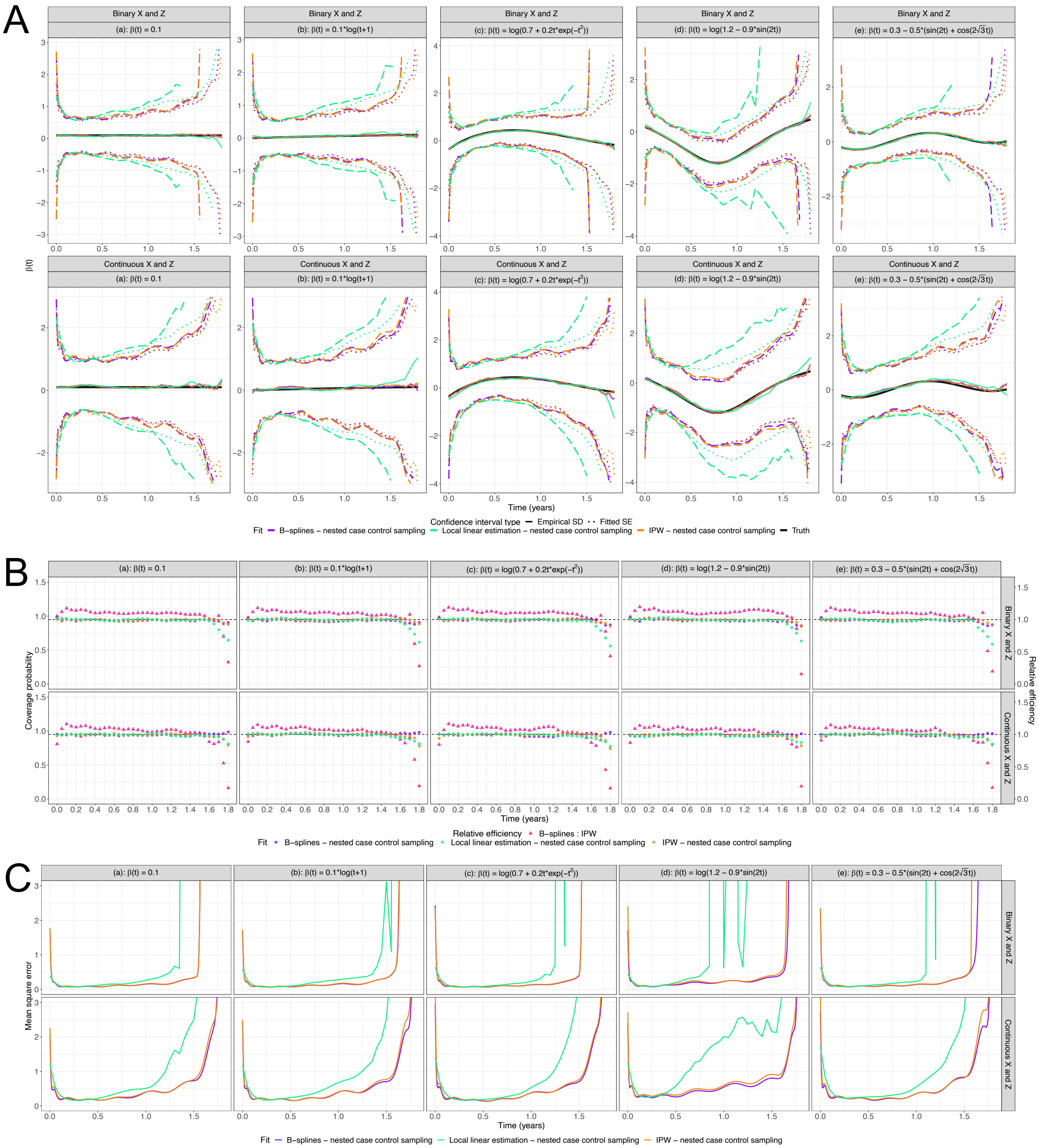

Figure 4.3: Simulation results for moderate correlation, nested case control sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

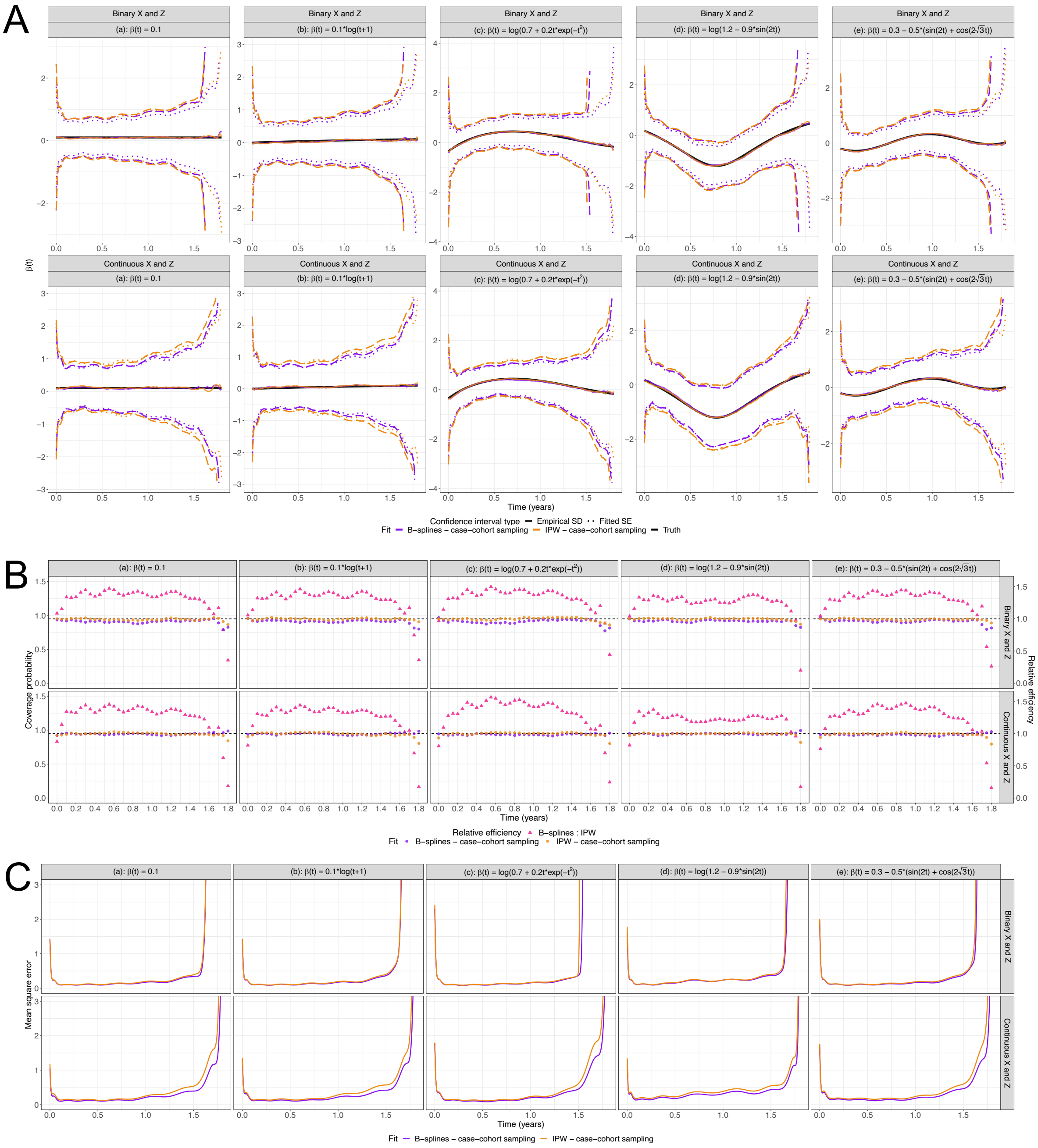

Figure 4.4: Simulation results for high correlation, case-cohort sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

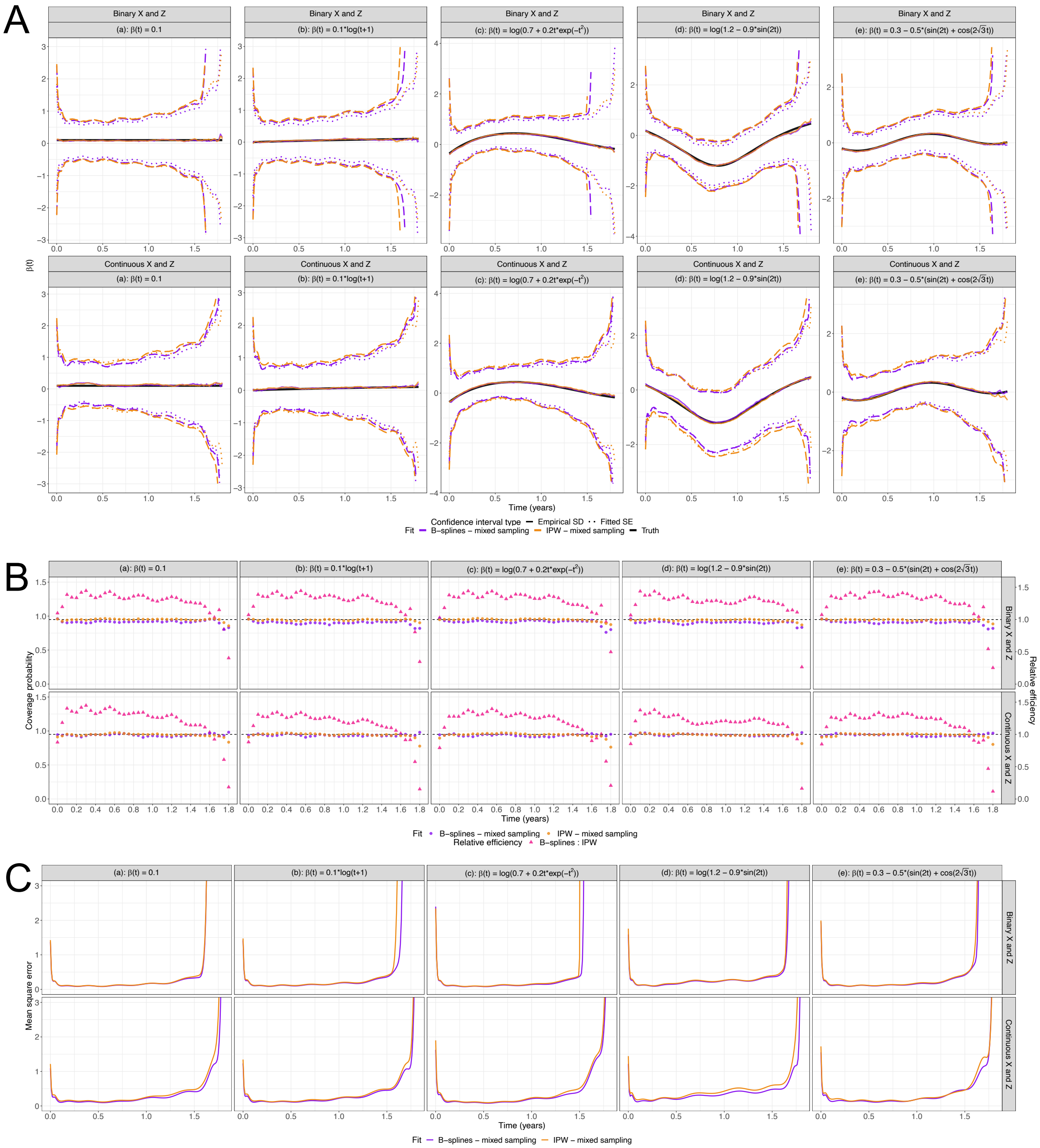

Figure 4.5: Simulation results for high correlation, mixed sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

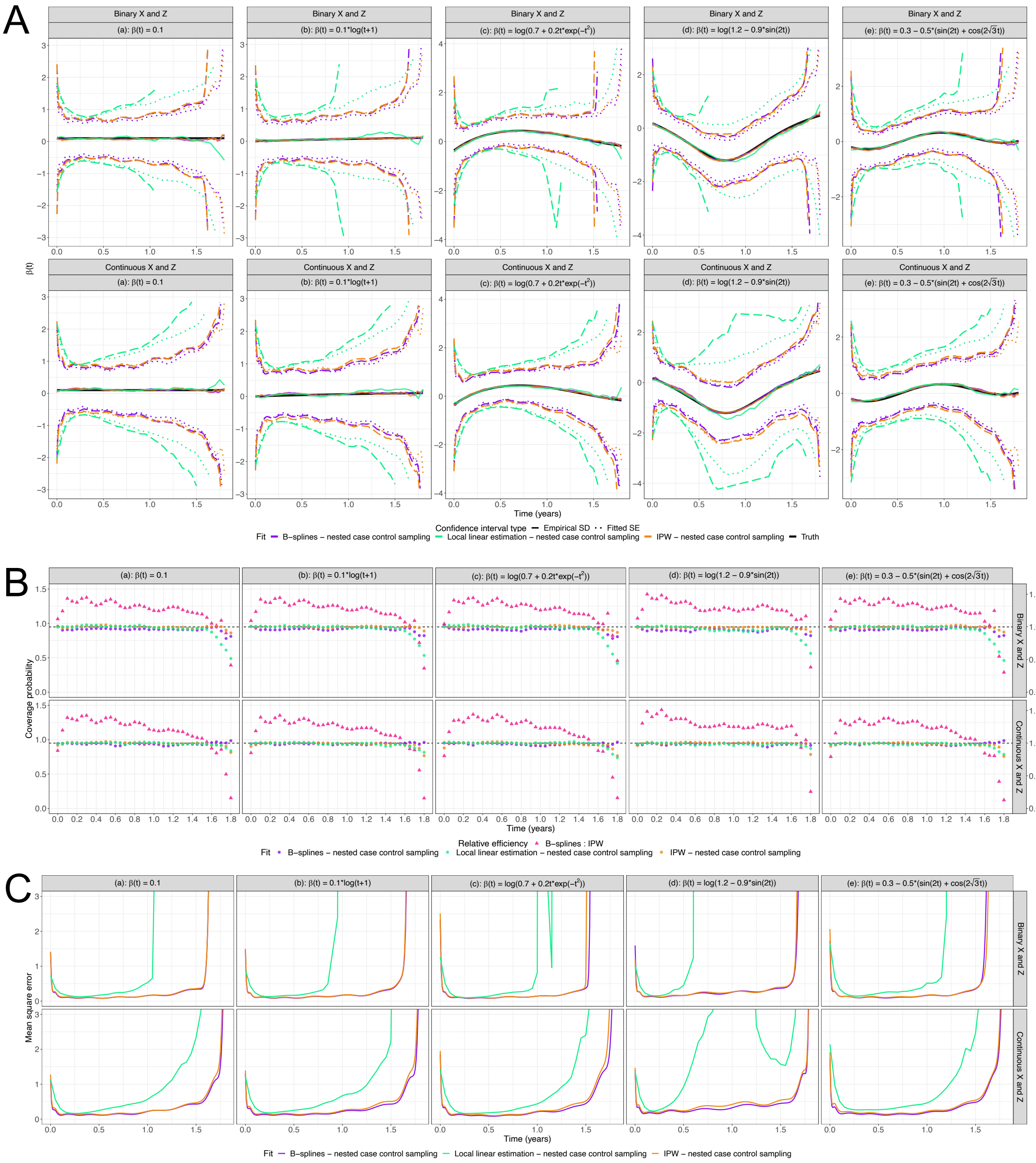

Figure 4.6: Simulation results for high correlation, nested case control sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

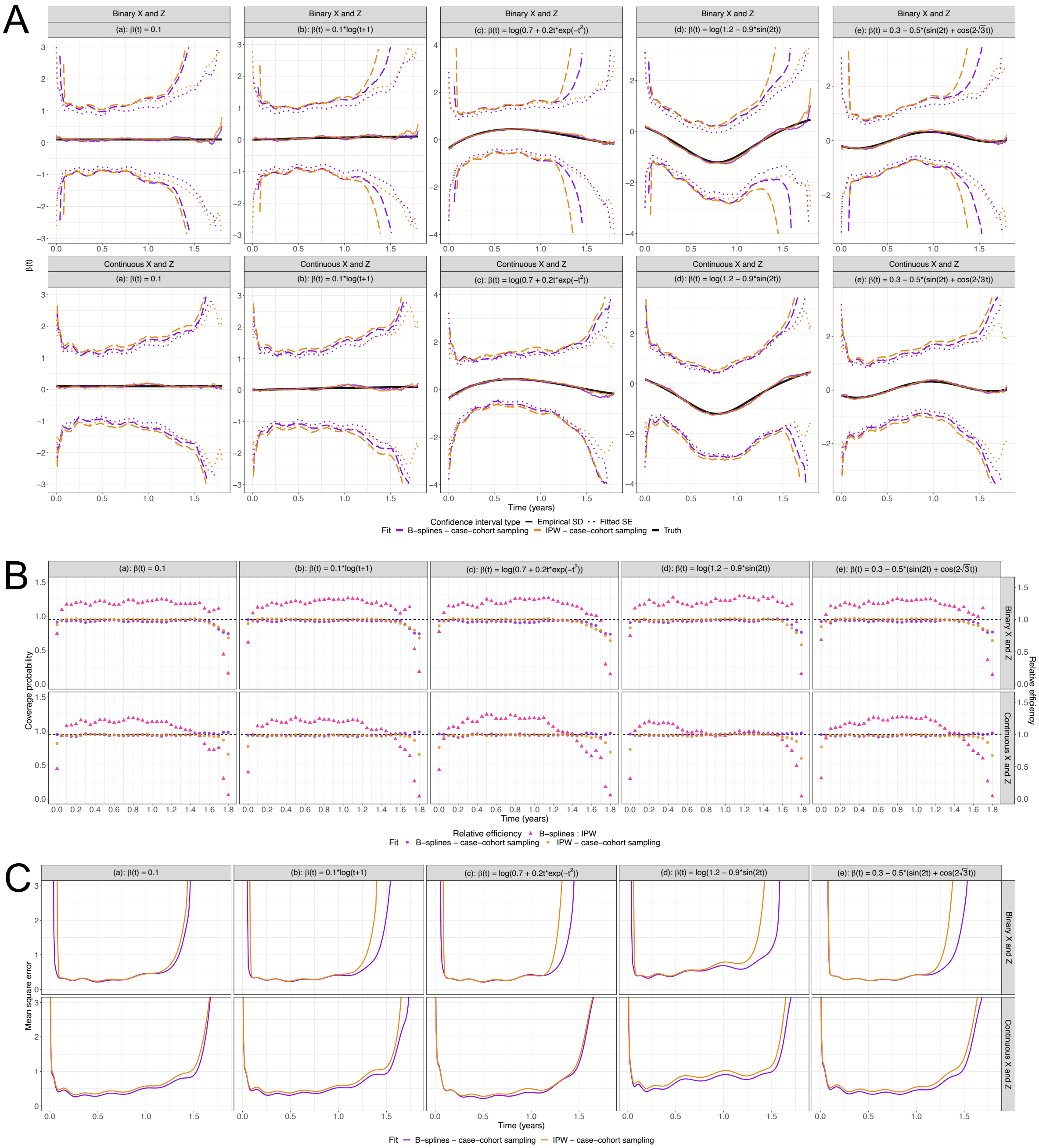

Figure 4.7: Simulation results for high correlation, rare disease, case-cohort sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

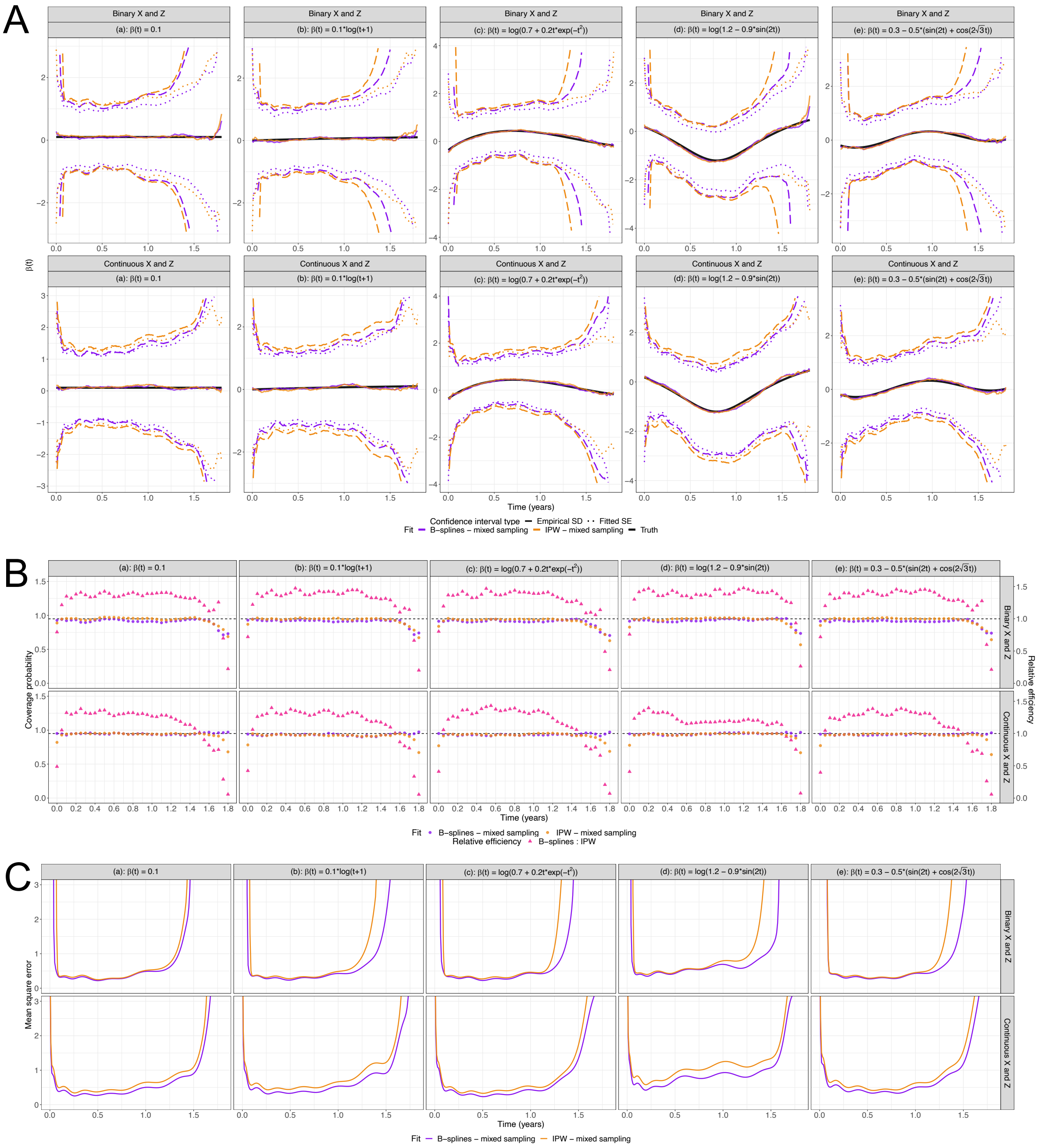

Figure 4.8: Simulation results for high correlation, rare disease, mixed sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

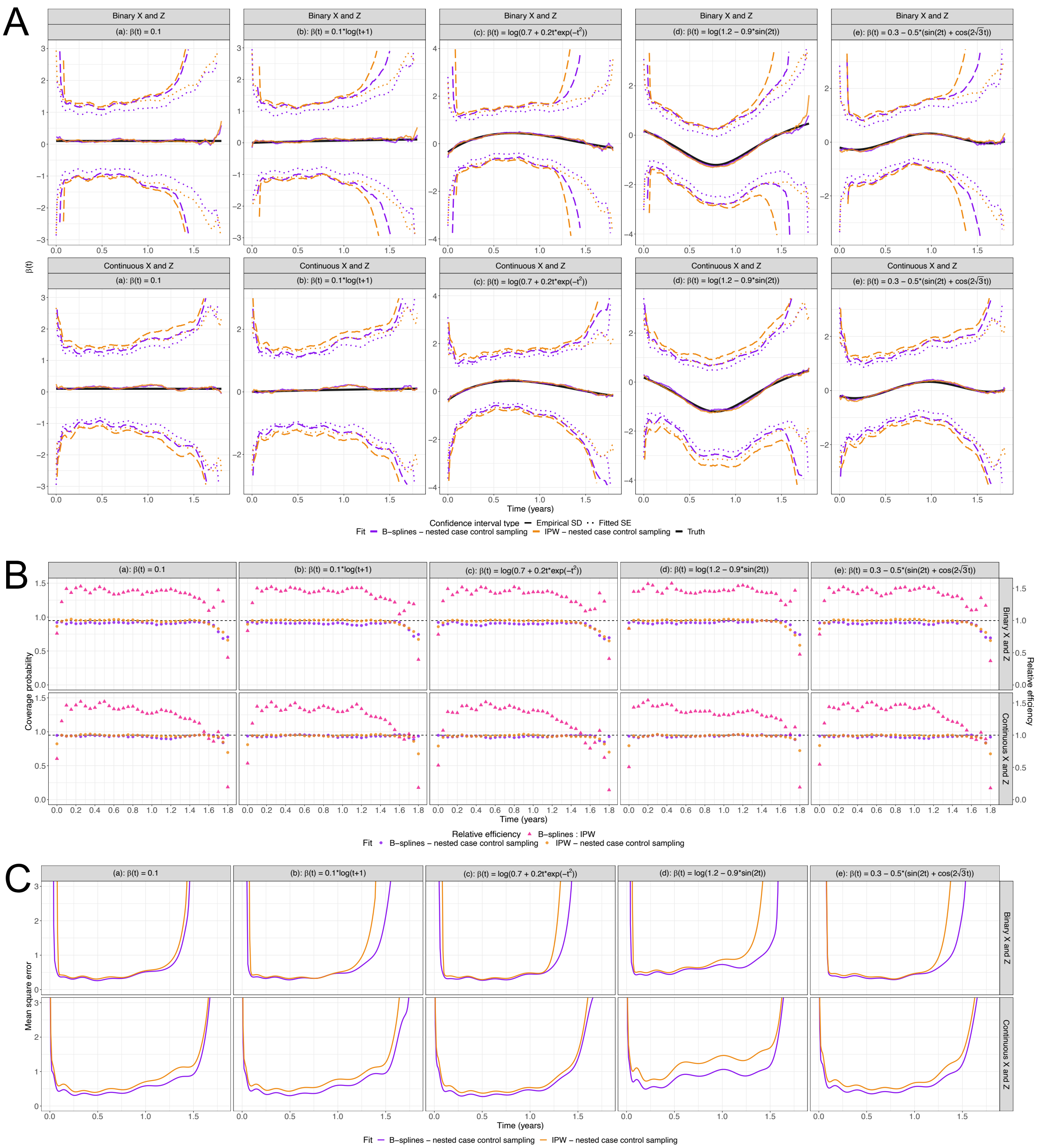

Figure 4.9: Simulation results for high correlation, rare disease, nested case control sampling, colored by fit type. (A) Estimated median curves and 95% confidence intervals, empirical and fitted (B) Coverage probability and relative efficiency (C) Mean square error.

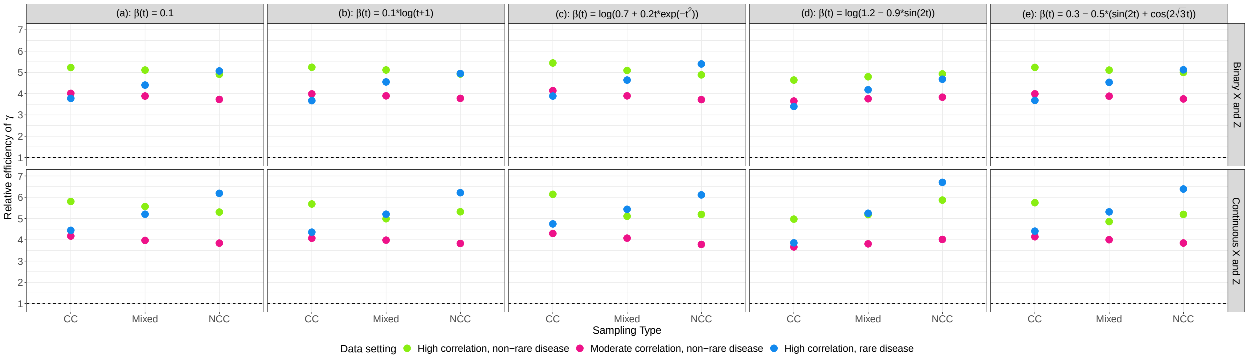

Figure 4.10: Simulation results for relative efficiency in γ estimation, by data setting (color) and sampling type (x-axis)

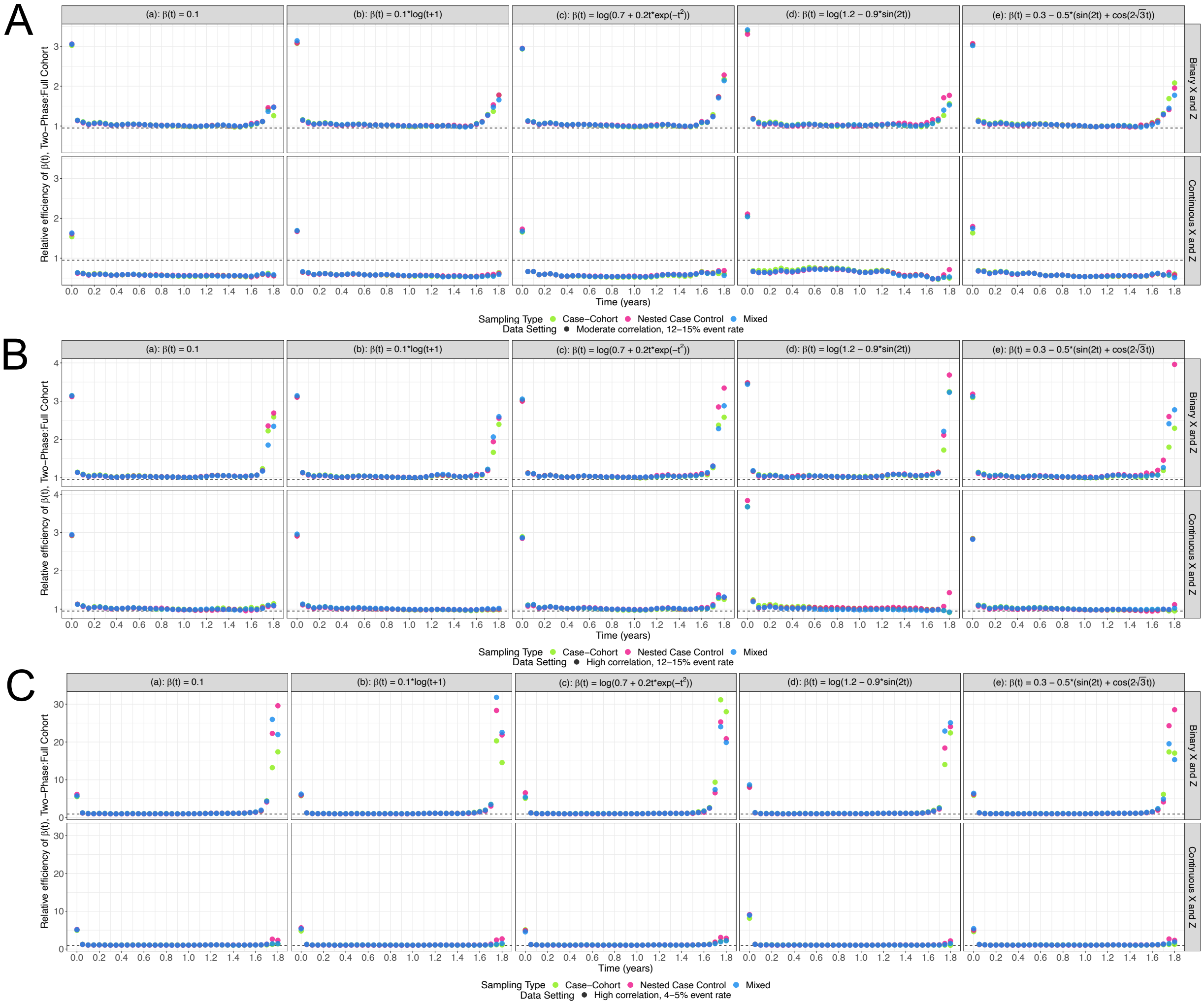

Figure 4.11: Simulation results for relative efficiency in $\beta(t)$ estimation between two-phase B-spline fit and full cohort fit, by data setting and sampling type. (A) Moderate correlation, $10-15\%$ event rate (B) High correlation, $10-15\%$ event rate (C) High correlation, $4-5\%$ event rate.

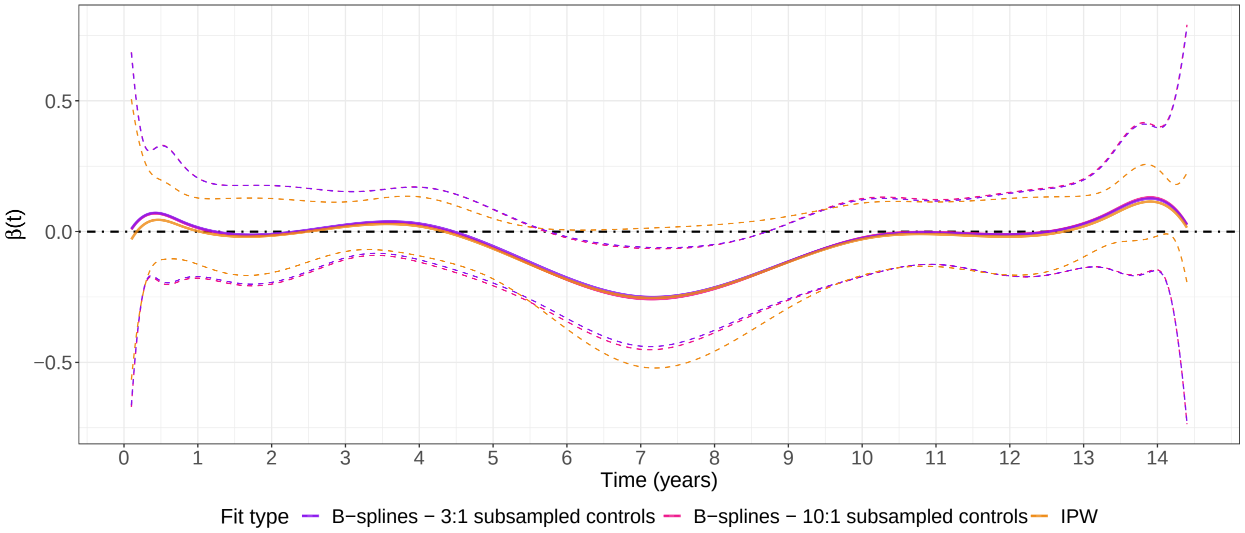

Figure 4.12: Resulting estimates of $\beta(t)$ and corresponding 95% confidence intervals fit to colorectal cancer data, for two-phase B-splines sub-sampled fits and IPW fit.

CHAPTER 5

Conclusion

This dissertation aims to provide methods for improving both inferential and computational efficiency when analyzing real-world data sets, particularly those coming from clinical trials or epidemiological studies. These large data sources provide opportunities for researchers to explore rare events, control for additional sources of confounding that may previously have been overlooked, and answer questions beyond those originally posed by the study due to the availability of additional covariates. However, existing methods cannot always scale to accommodate ever-larger amounts of data, increasing computational needs beyond feasibility, and it may also not be possible to collect complete information on every subject and variable of interest due to time, budget, or access constraints. We have presented several methods for handling different types of large or complex data scenarios, each with applications to cancer or chemotherapy outcomes.

In Chapter 2, we developed a dose toxo-equivalence model using a Bayesian meta-analysis of published clinical trials. Our model was equally efficient to the equivalent model fit to individual patient data in most settings, and was able to provide reasonable dosing guidance even under misspecification that may arise from the use of summarized versus complete data. The resources needed to perform this study-level meta-analysis were significantly less than those needed for an individual patient data analysis, both in data collection (effort to gather summary data from trials versus contacting each study author individually to request access to complete patient information) and in computation (number of rows in data becomes total number of studies, rather than total number of patients across all studies). Given that it is often difficult or impossible to collect complete patient information from previously run clinical trials, this approach allows researchers to leverage existing publicly available information to guide dosing decisions in clinical practice.

In Chapter 3, we compared existing methods for estimating time-varying effects in the Cox model. Particularly with the long follow-up times seen in large clinical and epidemiological studies, the assumption of constant covariate effects across time is more likely to be violated. The non-parametric estimation of time-varying effects allows researchers to accommodate these non-proportional hazards without making additional assumptions about the nature of these time-varying effects. We considered five different approaches, and found that their efficiency and computational needs differed based on tuning parameter selection as well as the complexity of the underlying effect being estimated. Our application to a data set that was both long ($> 40,000$ subjects) and wide ($> 2,400$ covariates) revealed which methods were able to accommodate the types of data that come from large epidemiological studies, particularly those involving EHR data. Ultimately, we found estimation via B-splines had the most consistent performance across scenarios and was the least sensitive to tuning parameter selection, providing a flexible option with which to estimate time-varying effects.

In Chapter 4, we again considered time-varying effects in the Cox model, but this time in the context of two-phase study designs. We developed an efficient semiparametric method that uses B-spline sieves to approximate the conditional density functions of expensive covariates given available inexpensive covariates, and showed that this method is able to match the efficiency of a model fit when expensive covariates are available for every subject from Phase I. By incorporating all information from Phase I, our method outperformed existing estimation methods that consider only subjects from Phase II with complete covariate information. Additionally, we proposed several methods for reducing the computational intensity of our approach when faced with particularly large Phase I samples, including subject sub-sampling, dimension reduction via PCA, and the rounding of covariates and event times to reduce the total number of unique values.

Altogether, this dissertation addresses several issues that may arise when working with data from large clinical or epidemiological studies. Our methods provide computationally feasible options for analysis that yield efficient estimators even in the face of incomplete or complex data. We believe that this work builds on and extends existing statistical methodologies in the face of the challenges posed by the increasing availability of large real-world data.

REFERENCES

Argyriou, A. A., Koltzenburg, M., Polychronopoulos, P., Papapetropoulos, S., and Kalofonos, H. P. (2008). Peripheral nerve damage associated with administration of taxanes in patients with cancer. *Critical Reviews in Oncology/Hematology*, 66(3):218–228.

Austin, P. C., Fang, J., and Lee, D. S. (2022). Using fractional polynomials and restricted cubic splines to model non-proportional hazards or time-varying covariate effects in the cox regression model. *Statistics in Medicine*, 41(3):612–624.

Berger, J. O., Liseo, B., Wolpert, R. L., et al. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., and Feldman, H. I. (2002). Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine*, 21(3):371–387.

Bjelakovic, G., Nikolova, D., Simonetti, R. G., and Gluud, C. (2004). Antioxidant supplements for prevention of gastrointestinal cancers: a systematic review and meta-analysis. *The Lancet*, 364(9441):1219–1228.

Blot, W. J., Li, J.-Y., Taylor, P. R., Guo, W., Dawsey, S., Wang, G.-Q., Yang, C. S., Zheng, S.-F., Gail, M., Li, G.-Y., et al. (1993). Nutrition intervention trials in linxian, china: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population. *JNCI: Journal of the National Cancer Institute*, 85(18):1483–1491.

Brunet, A., Sweeney, L. B., Sturgill, J. F., Chua, K. F., Greer, P. L., Lin, Y., Tran, H., Ross, S. E., Mostoslavsky, R., Cohen, H. Y., et al. (2004). Stress-dependent regulation of foxo transcription factors by the sirt1 deacetylase. *Science*, 303(5666):2011–2015.

Cai, T., Parast, L., and Ryan, L. (2010). Meta-analysis for rare events. *Statistics in Medicine*, 29(20):2078–2089.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111.

Cao, Y., Lin, H., Wu, T. Z., and Yu, Y. (2010). Penalized spline estimation for functional coefficient regression models. *Computational Statistics & Data Analysis*, 54(4):891–905.

Cates, C. J. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Medical Research Methodology*, 2(1):1.

Chandel, N. S. and Tuveson, D. A. (2014). The promise and perils of antioxidants for cancer patients. *New England Journal of Medicine*, 371(2):177–178.

Chen, Q., Zeng, D., Ibrahim, J. G., Akacha, M., and Schmidli, H. (2013). Estimating time-varying effects for overdispersed recurrent events data with treatment switching. *Biometrika*, 100(2):339–354.

Cole, B. F., Baron, J. A., Sandler, R. S., Haile, R. W., Ahnen, D. J., Bresalier, R. S., McKeown-Eyssen, G., Summers, R. W., Rothstein, R. I., Burke, C. A., et al. (2007). Folic acid for the prevention of colorectal adenomas: a randomized clinical trial. *JAMA*, 297(21):2351–2359.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

Debray, T. P., Moons, K. G., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H., Reitsma, J. B., and Group, G. M. R. (2015). Get real in individual participant data (ipd) meta-analysis: a review of the methodology. *Research Synthesis Methods*, 6(4):293–309.

DeNicola, G. M., Karreth, F. A., Humpton, T. J., Gopinathan, A., Wei, C., Frese, K., Mangal, D., Yu, K. H., Yeo, C. J., Calhoun, E. S., et al. (2011). Oncogene-induced nrf2 transcription promotes ros detoxification and tumorigenesis. *Nature*, 475(7354):106–109.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., and Crawford, D. C. (2010). Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121.

Gaziano, J. M., Glynn, R. J., Christen, W. G., Kurth, T., Belanger, C., MacFadyen, J., Bubes, V., Manson, J. E., Sesso, H. D., and Buring, J. E. (2009). Vitamins e and c in the prevention of prostate and total cancer in men: the physicians' health study ii randomized controlled trial. *JAMA*, 301(1):52–62.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Genestra, M. (2007). Oxyl radicals, redox-sensitive signalling cascades and antioxidants. *Cellular Signalling*, 19(9):1807–1819.

Group, A.-T. B. C. C. P. S. (1994). The effect of vitamin e and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine*, 330(15):1029–1035.

Harrell, F. E. et al. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer.

Harrell Jr, F. E., with contributions from Charles Dupont, and many others. (2020). *Hmisc: Harrell Miscellaneous*. R package version 4.4-1.

Hastie, T., Sleeper, L., and Tibshirani, R. (1992). Flexible covariate effects in the proportional hazards model. *Breast Cancer Research and Treatment*, 22(3):241–250.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hothorn, T. (2020). Most likely transformations: The mlt package. *Journal of Statistical Software*, 92(1):1–68.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293.

Kh, V. and Ryan, K. (2009). p53 and metabolism. *Nature Reviews Cancer*, 9(10):691–700.

Klein, E. and Thompson Jr, I. (2011). Tangen cm, et al. vitamin e and the risk of prostate cancer: the selenium and vitamin e cancer prevention trial (select). *JAMA*, 306:1549–1556.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, 90(429):78–94.

Kops, G. J., Dansen, T. B., Polderman, P. E., Saarloos, I., Wirtz, K. W., Coffer, P. J., Huang, T.-T., Bos, J. L., Medema, R. H., and Burgering, B. M. (2002). Forkhead transcription factor foxo3a protects quiescent cells from oxidative stress. *Nature*, 419(6904):316–321.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A tutorial with R, Jags, and Stan*. Academic Press.

Lambert, P. C., Sutton, A. J., Abrams, K. R., and Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, 55(1):86–94.

Lebovits, A. H., Strain, J. J., Messe, M. R., Schleifer, S. J., Tanaka, J. S., and Bhardwaj, S. (1990). Patient noncompliance with self-administered chemotherapy. *Cancer*, 65(1):17–22.

Liddell, F., McDonald, J., and Thomas, D. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *Journal of the Royal Statistical Society: Series A (General)*, 140(4):469–483.

Lin, D.-Y. and Zeng, D. (2010a). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):60–66.

Lin, D.-Y. and Zeng, D. (2010b). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332.

Liu, M., Lu, W., Shore, R. E., and Zeleniuch-Jacquotte, A. (2010). Cox regression model with time-varying coefficients in nested case–control studies. *Biostatistics*, 11(4):693–706.

Lorentz, G. G. (2013). *Bernstein polynomials*. American Mathematical Society.

Luo, C., Islam, M., Sheils, N. E., Buresh, J., Reps, J., Schuemie, M. J., Ryan, P. B., Edmondson, M., Duan, R., Tong, J., et al. (2022). Dlmm as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nature Communications*, 13(1):1–10.

Lutz, W., Lowry, J., Kopta, S. M., Einstein, D. A., and Howard, K. I. (2001). Prediction of dose–response relations based on patient characteristics. *Journal of Clinical Psychology*, 57(7):889–900.

Malloy, E. J., Spiegelman, D., and Eisen, E. A. (2009). Comparing measures of model selection for penalized splines in cox models. *Computational Statistics & Data Analysis*, 53(7):2605–2616.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, pages 115–117. Chapman and Hall, second edition.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Olkin, I. and Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, pages 317–322.

Omenn, G. S., Goodman, G. E., Thornquist, M. D., Balmes, J., Cullen, M. R., Glass, A., Keogh, J. P., Meyskens Jr, F. L., Valanis, B., Williams Jr, J. H., et al. (1996). Risk factors for lung cancer and for intervention effects in caret, the beta-carotene and retinol efficacy trial. *JNCI: Journal of the National Cancer Institute*, 88(21):1550–1559.

Osman, M. and Ghosh, S. K. (2012). Nonparametric regression models for right-censored data using bernstein polynomials. *Computational Statistics & Data Analysis*, 56(3):559–573.

Pais, R. and Dumitraşcu, D. (2013). Do antioxidants prevent colorectal cancer? a meta-analysis. *Romanian journal of internal medicine= Revue roumaine de medecine interne*, 51(3-4):152–163.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in r. *BMC Medical Research Methodology*, 19(1):1–16.

Piskounova, E., Agathocleous, M., Murphy, M. M., Hu, Z., Huddlestun, S. E., Zhao, Z., Leitch, A. M., Johnson, T. M., DeBerardinis, R. J., and Morrison, S. J. (2015). Oxidative stress inhibits distant metastasis by human melanoma cells. *Nature*, 527(7577):186–191.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizos, C. V. and Elisaf, M. S. (2013). Metformin and cancer. *European Journal of Pharmacology*, 705(1-3):96–108.

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., and Masys, D. R. (2008). Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Saarela, O., Kulathinal, S., Arjas, E., and Läärä, E. (2008). Nested case–control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in Medicine*, 27(28):5991–6008.

Salem, N. and Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, 163:292–299.

Samuelsen, S. O. (1997). A psudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394.

Severini, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, 85(3):507–522.

Sigworth, E. A., Rubinstein, S. M., Chaugai, S., Rivera, D. R., Walker, P. D., Chen, Q., and Warner, J. L. (2022). Development of a bayesian toxo-equivalence model between docetaxel and paclitaxel. *iScience*, 25(4):104045.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85(412):941–949.

Steinberg, K., Smith, S., Stroup, D., Olkin, I., Lee, N., Williamson, G., and Thacker, S. (1997). Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *American Journal of Epidemiology*, 145(10):917–925.

Støer, N. C. and Samuelsen, S. O. (2012). Comparison of estimators in nested case–control studies with multiple outcomes. *Lifetime Data Analysis*, 18(3):261–283.

Støer, N. C. and Samuelsen, S. O. (2016). multiplencc: Inverse probability weighting of nested case-control data. *R Journal*, 8(2):5.

Su, Y.-S. and Yajima, M. (2020). *R2jags: Using R to Run 'JAGS'*. R package version 0.6-1.

Tao, R., Zeng, D., and Lin, D.-Y. (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association*, 112(520):1468–1476.

Tao, R., Zeng, D., and Lin, D.-Y. (2020). Optimal designs of two-phase studies. *Journal of the American Statistical Association*, 115(532):1946–1959.

Tenbusch, A. (1997). Nonparametric curve estimation with bernstein estimates. *Metrika*, 45(1):1–30.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. and Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer.

Tian, L., Zucker, D., and Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183.

Verweij, P. J. and van Houwelingen, H. C. (1995). Time-dependent effects of fixed covariates in cox regression. *Biometrics*, pages 1550–1556.

Warner, J. L., Cowan, A. J., Hall, A. C., and Yang, P. C. (2015). Hemonc. org: A collaborative online knowledge platform for oncology professionals. *Journal of Oncology Practice*, 11(3):e336–e350.

Weyer, V. and Binder, H. (2015). A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. *BMC Bioinformatics*, 16(1):1–12.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. John Wiley & Sons.

Wu, Y., Warner, J. L., Wang, L., Jiang, M., Xu, J., Chen, Q., Nian, H., Dai, Q., Du, X., Yang, P., et al. (2019). Discovery of noncancer drug effects on survival in electronic health records of patients with cancer: a new paradigm for drug repurposing. *JCO Clinical Cancer Informatics*, 3:1–9.

Xu, H., Aldrich, M. C., Chen, Q., Liu, H., Peterson, N. B., Dai, Q., Levy, M., Shah, A., Han, X., Ruan, X., et al. (2015). Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*, 22(1):179–191.

Zeng, D. and Lin, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109(505):371–383.

Zeng, D. and Lin, D.-Y. (2015). On random-effects meta-analysis. *Biometrika*, 102(2):281–294.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1):329–353.