Causal Inference Methods and Intermediate Endpoints in Randomized Clinical Trials

by

Emily Kate Roberts

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Michael Elliott, Co-Chair
Professor Jeremy M.G. Taylor, Co-Chair
Assistant Professor Walter Dempsey
Associate Professor Ben Hansen

Emily Kate Roberts
ekrobe@umich.edu
ORCID iD: 0000-0002-5838-9691

DEDICATION

To my parents, the truest constants in my life during this PhD journey.

ACKNOWLEDGMENTS

First and foremost, the two people who I must acknowledge are my parents. Thank you for supporting me, loving me, and keeping me grounded. Thank you for being the place I could be vulnerable during the ups and downs of this PhD. I am blessed to have parents who were always diligent and supportive in the face of challenges. I am endlessly grateful to have you as my parents and for your sacrifices that made me who I am today. Thank you to my brother and family for your support, including the joy and giggles from my nephew Braxton. I never imagined I would write part of my dissertation at my childhood home during a global pandemic, but I am glad I was able to, and I am happy I will be coming back to Iowa for the next chapter in my journey.

Next, thank you to my dissertation co-advisor and first mentor at Michigan through the Cancer Training grant, Jeremy Taylor. You have always been generous with your time and expertise over the years, and I am grateful to have had my graduate training with you as a mentor. Thank you also to my co-advisor Michael Elliott, who has provided knowledge and guidance on my projects. Thank you also to Walter Dempsey and Ben Hansen for serving on my dissertation committee and providing valuable feedback and support. I am lucky that I have had the privilege of working with several mentors during my graduate studies: Lili Zhao, Phil Boonstra, Bhramar Mukherjee, Belinda Needham, Xu Shi, and Matthew Schipper, among others. Each of you has played a special role during my training and enhanced my preparation as a statistician. I also appreciate the collaborators I have had the honor to work with at the University and elsewhere, including the departments of pathology, radiation oncology, microbiology and immunology, Precision Health Data Science

Center, and others in the Cancer Center. I am grateful for the funding support I received during my PhD: the McElroy Foundation, the School of Public Health Regents Fellowship, the National Science Foundation Graduate Research Fellowship, the National Institutes of Health Cancer Research Training Grant, the Telomere Research Network Pilot Grant, and the Rackham Predoctoral Fellowship Program. I know I have been blessed with many opportunities to attend workshops and conferences thanks to support from many sources.

In addition to the resources and opportunities at Michigan, this dissertation was possible because of the support of many individuals and experiences that shaped me personally and professionally. Thank you to the educators who believed in me over the years. These include those at summer biostatistics undergraduate research programs that confirmed my passions for this field, and at Coe College such as mentors Drs. Jonathan White, Nükhet Yarbrough, Dan Lehn, Karla Steffens-Moran, and others. Thank you to those from my grade school years, like Nancy and school nurses I visited each day from kindergarten to high school at lunchtime – thanks to these individuals and my parents, I never feared type one diabetes would interfere with my academic pursuits. Instead, it is my motivator to contribute to meaningful medical research.

I am grateful for my time spent in the biostatistics department and the opportunity to make Ann Arbor my second home during this journey; this included learning from brilliant mentors and classmates and growing through struggle and triumph. Working with individuals through organizations such as Girls Who Code and Statistics in the Community, STATCOM, enhanced my training by connecting my academic and personal passions. Thank you to my peers who have been by my side during an eventful PhD. In particular, thank you Sarah, Brooke, Andrea, Elizabeth, Madeline, and Aarohee, who have become dear friends during our doctoral studies. I am grateful for Tian, Lauren, Krithika, and others for serving as mentors while our graduate careers overlapped. I am thankful for all of my office-mates, cohort-mates, and peers from various

lab groups that have cheered me on and served as role models to me. Sharing this journey with each of you is a privilege, and many times I needed your listening ears and kind support to make it through. Precious friendships outside of school kept me balanced and connected to life outside of the PhD as well. Thanks especially to my dearest Dance Revolution family. Salsa dancing became an important and joyful outlet for me during graduate school, but the friends I have made along the way have been an unexpected blessing that I cannot imagine this chapter in my life without.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGMENTS iii

LIST OF FIGURES ix

LIST OF TABLES xi

LIST OF APPENDICES xiii

ABSTRACT xiv

CHAPTER

I Introduction 1

II Incorporating Baseline Covariates to Validate Surrogate Endpoints with a Constant Biomarker Under Control Arm 8

2.1 Introduction 8
2.2 The Model 12
2.2.1 Assumptions 13
2.2.2 Baseline Covariates 14
2.2.3 Surrogacy Validation 15
2.2.4 Subgroups in Surrogacy and Treatment Effects 17
2.2.5 Conditional Independence 17
2.2.6 Design Considerations and Defining the Endpoint 18
2.3 Bayesian Methods 20
2.3.1 Imputation-Estimation Algorithms 20
2.3.2 Observed Data Algorithm 21
2.4 Simulation Studies 23
2.4.1 Simulation Results 25
2.4.2 Marginal Estimates 26
2.4.3 Sensitivity to Distributional Assumptions 27
2.4.4 Subgroup Analysis 27
2.5 Duchenne Muscular Dystrophy Data Example 28
2.5.1 Data Example Results 29

2.6 Discussion . . . 30
2.7 Publication . . . 32
2.8 Figures . . . 33

III Solutions for Surrogacy Validation with Longitudinal Outcomes for a Gene Therapy 36

3.1 Introduction . . . 36
3.2 The Model . . . 39
3.2.1 Repeated Measurements in a Standard Trial Design . . . 39
3.2.2 Pre-treatment Observations of T . . . 41
3.2.3 Delayed-Start of Treatment Design . . . 42
3.2.4 Conditional Independence . . . 43
3.3 Surrogacy Validation . . . 44
3.4 Estimation Methods for Standard Trial Design . . . 46
3.4.1 Counterfactual Imputation Methods . . . 47
3.4.2 Observed Data Methods with Random Effects . . . 48
3.4.3 Observed Data Methods Integrating Out Random Effects . . . 49
3.4.4 Methods Simplifying the Repeated Trial Outcome . . . 50
3.5 Simulation Studies . . . 51
3.5.1 Results Comparing Methods and Models with Random Intercepts . . . 51
3.5.2 Results Comparing Assumptions with Random Slopes . . . 54
3.5.3 Results for Pretreatment Measures . . . 55
3.6 Motivating Data Example . . . 55
3.7 Discussion and Future Directions . . . 57
3.8 Data Availability and Software . . . 59
3.9 Tables . . . 60

IV Surrogacy Validation for Time-to-Event Outcomes with Illness-Death Frailty Models 62

4.1 Introduction . . . 62
4.2 Illness-Death Approach . . . 66
4.2.1 Defining Causal Quantities Based on Hazards and Frailty Models . . . 68
4.2.2 Identifiability and Sensitivity Analysis . . . 73
4.3 Likelihood and Estimation . . . 73
4.3.1 Likelihood Contributions . . . 73
4.3.2 Bayesian Estimation . . . 75
4.4 CEP Quantities . . . 77
4.4.1 Valid Surrogates under an Illness-Death CEP Curve . . . 80
4.5 Simulation Study . . . 82
4.5.1 Simulation Set-up . . . 82
4.5.2 Simulation Results . . . 84
4.6 Data Example . . . 86
4.6.1 Conventional Models . . . 89
4.6.2 Surrogacy Evaluation . . . 90
4.7 Discussion and Future Work . . . 93
4.8 Tables . . . 97

V Conclusion . . . 102

APPENDICES 115

BIBLIOGRAPHY 156

LIST OF FIGURES

FIGURE

I.1 Example CEP Plot demonstrating valid and invalid potential surrogates based on the validation quantities. γ_0 can be interpreted as the intercept and γ_1 as the slope of the CEP curves in this case. 4

2.1 CEP plot showing the conditional functions of γ_0 and γ_1 across possible values of $S(1)$. 29
2.2 Simulation results for the two definitions of the endpoint 34

3.1 Counterfactual data for the delayed-start treatment design. 43
3.2 Results of the CEP plot for the delayed-treatment Muscular Dystrophy trial simulated data. 57

4.1 Counterfactual illness-death models for baseline, illness (S), and death (T). The potential pathways are labeled with the gap time and corresponding transition intensity for each treatment arm. 66
4.2 Relationships between the parameters and data in the proposed causal illness-death model. 77
4.3 Example of an estimated CEP curve, conditional on frailties, for a single simulated dataset where we assume all values of κ_{jk}^z are fixed. 86
4.4 Kaplan Meier curves for the intermediate and true outcome and the Kaplan Meier curve for the transition from S to T among those who experienced S for the prostate cancer trial. 88
4.5 Cumulative incidence curves for the two treatment groups in the prostate cancer clinical trial. 89
4.6 Counterfactual Illness-Death Models for baseline, illness (S), and death (T) with the number of individuals experiencing the events in each transition for the prostate cancer trial. 90
4.7 Causal effect predictiveness plot for the motivating prostate cancer trial dataset. 92
4.8 Causal effect predictiveness plot for the motivating prostate cancer trial dataset. 93

5.1 Potential associations between the treatment effects, Δ_S and Δ_T, for different treatment effects relevant for futility analysis. 111

4.1 Simulation results and sensitivity analysis of data example results over different values of θ_T using the observed data method 121
4.2 Simulation results and sensitivity analysis of data example results over different values of θ_T using the imputation method 122

14.1 Scenarios 1-4: Null, perfect, and partial surrogates under the illness-death formulation. 144
14.2 Scenarios 5-8: Partial and non-surrogates under the illness-death formulation. 145
14.3 CEP curve when $\rho_S = \rho_T = 0$ versus when $\rho_S = \rho_T = 0.95$. 146
14.4 CEP curves comparing changing $\theta_{23}^z = -1, 0, 1$. 147

LIST OF TABLES

TABLE

2.1 Generative parameter values for the six scenarios to compare definition of the endpoint and using baseline covariates. . . . 33
2.2 Simulation results demonstrating effect of estimating subgroups. . . . 33
2.3 Simulated muscular dystrophy estimates of treatment effect and marginal surrogacy quantities. . . . 35

3.1 Simulation results of random intercept models comparing different assumptions and models. . . . 60
3.2 Simulation results of random slope models over time. . . . 61
3.3 Simulation results of random slope models comparing trial designs. . . . 61

4.1 Eight possible scenarios of which pathways in the illness-death models exhibit treatment effects based on the causal hazards including an intuitive notion of whether S is a good surrogate for T. . . . 97
4.2 Simulation results from illness-death models and estimated validation quantities with κ_{jk} parameters fixed. . . . 98
4.3 Simulation results from illness-death models and parameter estimates with κ_{jk} parameters fixed. . . . 99
4.4 Simulation results from illness-death models and regression coefficients with κ_{jk} parameters fixed. . . . 100
4.5 Parameter estimates for the prostate cancer data example. . . . 101

3.1 Simulation results demonstrating different definitions of the endpoint and different generating parameter values . . . 120

5.1 Simulation results demonstrating effect of non-normal distributions when fitting conditional models that assume normality. . . . 123

13.1 Simulation results of random intercept models comparing different sample sizes for the observed data algorithm . . . 139
13.2 Simulation results of random intercept models for misspecified models . . . 140
13.3 Simulation results of random slopes models with varying sample sizes . . . 140
13.4 Simulation results of random slopes models for misspecified models where the true generating model includes a non-linear (quadratic) term . . . 141

14.1 Eight possible scenarios of which pathways in the illness death models exhibit treatment effects based on the causal hazards. . . . 143

17.1 Simulation results from illness-death models and estimated validation quantities with scale parameters fixed. 153
17.2 Simulation results from illness-death models of regression coefficients when scale parameters are fixed. 153

LIST OF APPENDICES

A Imputation Algorithm Details **115**

B The Four Trivariate Normal Distributions and Corresponding Surrogacy Quantities **117**

C Tables for Simulation Results Demonstrating Different Definitions of the Endpoint and Different Generating Parameter Values **119**

D Simulation Results and Sensitivity Analysis of Data Example Results **121**

E Simulation Results Demonstrating Effect of Non-normal Distributions **123**

F Generative Parameter Values for Plausible Clinical Trial Data Example **124**

G Counterfactual Imputation Details **125**

H Derivations for Random Slopes Details **128**

I Fisher Approximation for Correlation Details **131**

J Delayed-Start Treatment Design Details **133**

K Generating Proper Covariance Matrices **136**

L Optimization Method Details **138**

MSimulation Results for Sensitivity Analyses **139**

N Defining Ideal CEP Curves for Time-to-Event Data **142**

O Prentice Approach Formulation and Relation to Proposed Illness-Death Method . . . **148**

P Likelihood Contributions for Illness-Death Model Parameters **150**

Q Simulation Results for Illness-Death Model Parameters **152**

R Rshiny App for Illness-Death Model Parameters **154**

ABSTRACT

In clinical research and randomized clinical trials, intermediate endpoints can serve several purposes. It is possible that an intermediate marker may serve as a surrogate S for a true clinical outcome of interest T with the goal of making the trial run more efficiently or cost-effectively. Rigorous assessment as to whether a proposed surrogate endpoint is valid is challenging, however.

Chapter II extends causal inference approaches to validate a candidate surrogate outcome using potential outcomes. Using the principal surrogacy criteria, we incorporate baseline covariates in the setting of normally-distributed endpoints. In particular, our setting of interest allows us to assume the surrogate under the placebo, $S(0)$, is zero-valued. We develop methods to incorporate conditional independence and other modeling assumptions and explore their impact on the assessment of surrogacy. We demonstrate our approach via simulation of data that mimics an ongoing study of a muscular dystrophy gene therapy.

Chapter III also considers the motivating clinical trial for muscular dystrophy, whereas now the true outcomes $T(0), T(1)$ are measured longitudinally. We develop a mixed model approach that can potentially gain estimation efficiency. Further, it may be possible to measure additional T and S outcomes in a delayed treatment start or cross-over trial design. In this situation, subjects who are first administered the placebo may be given the gene therapy at a later time. This chapter addresses models and metrics for validation in such a trial. We also consider how to define the quantities for validation such that they may depend on time.

In Chapter IV, we extend these ideas to the surrogate validation framework with time-to-event

data. We develop a method that incorporates the censoring and semi-competing risk structure that is often encountered with multiple survival endpoints. We consider novel ways to define the parameters measuring the association between outcomes and relevant principal strata using a illness-death framework. We model conditional hazards while maintaining a valid causal interpretation by viewing this through the lens of a causal multi-state model. Finally, we apply our proposed methods to a prostate cancer randomized clinical trial.

CHAPTER I

Introduction

In randomized clinical trials, intermediate endpoints can serve multiple important purposes. It is possible that an intermediate marker may serve as a surrogate for a true clinical outcome of interest with the goal of making the trial run more efficiently or cost-effectively. Popular examples of potential surrogate endpoints include CD4 blood counts for HIV mortality and immune responses for vaccine efficacy. Rigorous assessment as to whether a proposed surrogate endpoint is valid is challenging, however. This dissertation extends causal inference approaches to validate a candidate surrogate outcome using potential outcomes. We provide methods for a variety of types of endpoints. Specifically, we consider when the outcomes are either normally-distributed, longitudinally measured, or time-to-event endpoints.

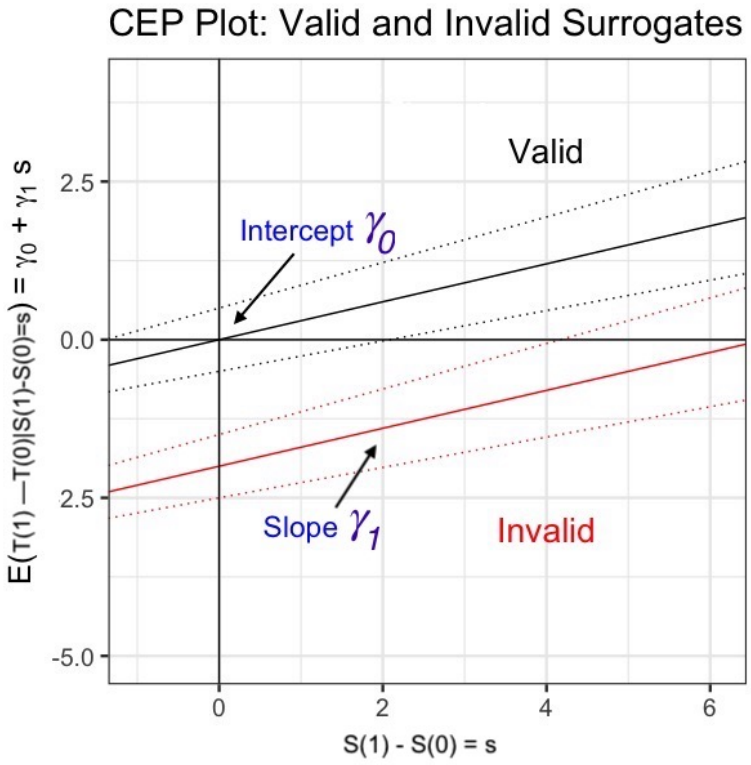
Prentice's landmark paper proposed two criteria to evaluate statistical surrogates in a single trial setting (1989). Based on associations between the surrogate and treatment and the surrogate and true outcome, the criteria require conditional independence between the treatment and outcome after adjusting for the surrogate value. These principles can be difficult to achieve and statistically validate in practice. While the method is applicable to a variety of endpoints for a single trial, it relies on conditioning on the observed value of S, which undesirably leads to a non-causal interpretation. Since the surrogate biomarker is measured after treatment assignment, adjusting for the surrogate distorts the causal pathway and interpretation of the treatment effect in a regression

model. Rubin's causal model proposes a framework where a potential outcome is the outcome that would have been observed under the opposite treatment (Rubin 1974; Little and Rubin 2000). However, the glaring inability to directly observe both outcomes and subsequent missing data problem has been deemed the "fundamental problem of causal inference."

More recent frameworks to determine if a surrogate marker is appropriate for use in a future trial can be broadly grouped into the causal effects and causal association paradigms (Joffe and Greene, 2009). The methods considered in this dissertation are based on principal stratification, a framework that was proposed as a solution to maintain a causal interpretation while properly incorporating the potential intermediate outcomes under both treatments (Frangakis and Rubin 2002). This work builds on the methods of Frangakis and Rubin and the corresponding Causal Effect Predictiveness (CEP) curve proposed by Gilbert and Hudgens (2008). Principal surrogacy can be used to assess a surrogate endpoint S for a true outcome T where $S(z)$ and $T(z)$ refer to the endpoint values had the treatment, possibly counter-factually, been assigned to level z. Since the potential outcomes are hypothetically determined prior to randomization, surrogates can be evaluated based on causal effects that are defined within principal strata based on $(S(0), S(1))$. In a simple case where S and T are Gaussian outcomes measured at one time point, surrogates can then be evaluated based on principal causal effects defined on the distribution of $T(1) - T(0)$ conditional on principal strata defined by $S(1) - S(0)$. In this case, the CEP curve for validation is based on $E(T(1) - T(0)|S(1) - S(0) = s)$. Gilbert and Hudgens elaborated on causal necessity, as proposed by Frangakis and Rubin, by terming average causal necessity and average causal sufficiency (2008). These conditions require that there be no average effect of the treatment on the true outcome in the strata where there is no average effect on the surrogate, and similarly that there exists an average treatment effect on the true outcome in the strata where is there an average effect on the surrogate.

The CEP curve is a visualization of both requirements across the value of the surrogate endpoints $s = (s_1, s_0)$. Broadly speaking, desired features of a surrogate are for this quantity to be close to zero when $s = 0$, and far from zero when s is not equal to zero. Figure I.1 graphically shows these concepts since we can write $E(T(1) - T(0)|S(1) - S(0) = s)$ as being equal to a function of two key parameters $\gamma_0 + \gamma_1 s$. In this setting, γ_0 corresponds to the intercept of the curve and γ_1 the slope. We see that the CEP curve in black labeled as "valid" goes through the origin and has a positive slope. Conversely, the red curve denoted as "invalid" fails to meet the criterion of going through the origin (corresponding to $\gamma_0 \neq 0$). Quantifying if a surrogate is valid will also depend on the uncertainty estimate around the curve. We build upon this CEP curve and principal stratification framework for different endpoint types and under different assumptions to aid in validation.

Figure I.1: Example CEP Plot demonstrating valid and invalid potential surrogates based on the validation quantities. γ_0 can be interpreted as the intercept and γ_1 as the slope of the CEP curves in this case.



Essentially we propose a series of statistical models and extract quantities in order to define specific estimands for validation. These quantities have been deemed appropriate to determine whether a surrogate endpoint is valid. Chapter II evaluates the principal stratification method performance under several model assumptions that are tailored to the motivating trial for muscular dystrophy. Specifically in this trial of a new gene transfer therapy, patients receive a micro-dystrophin transgene to produce the micro-dystrophin protein. The potential surrogate S is micro-dystrophin expression, and the primary outcome of interest T is a functional score on a continuous scale. In our trial setting, patients do not produce significant amounts of micro-dystrophin protein at baseline, so it can be assumed that the value of the surrogate under placebo, $S(0)$ is approximately equal to 0. This is an assumption that also naturally arises in vaccine efficacy trials and has

been termed the constant biomarker assumption (Gilbert and Hudgens, 2008).

Using the principal surrogacy criteria, we utilize the joint conditional distribution of the potential outcomes T, given the potential outcomes S. We build upon previous models of the joint distribution of potential outcomes $S(0),\ S(1),\ T(0),\ T(1)$ proposed by Conlon et al. (2014) by exploring trial and modeling considerations when only three of these endpoints vary, as is applicable to our motivating trial. Modeling the joint distribution of potential outcomes results in non-identified correlation parameters that must be addressed during estimation. We propose techniques and design considerations with the goal of achieving gains in estimation efficiency. One novel contribution is the motivation that we incorporate baseline covariates in the setting. Since muscular growth and deterioration due to the disease have major impact on physical movement during childhood, both baseline ambulatory ability and age are important to take into consideration. We develop methods to incorporate conditional independence and other modeling assumptions and explore their impact on the assessment of surrogacy. We then compare the estimation properties of a fully Bayesian imputation method using Markov Chain Monte Carlo to an algorithm using the observed data only. Within the simulation studies, we explore the impact of different prior distributions on non-identified parameters. We demonstrate our approach via simulation of data that mimics the ongoing muscular dystrophy study of the gene therapy.

Chapter III accommodates trials where the outcomes are measured longitudinally. In the same motivating trial of a gene therapy for muscular dystrophy patients, patients are followed over time, and subjects who are first administered the placebo may be given the gene therapy mid-trial. This chapter addresses models and metrics for surrogacy validation in such a trial. We propose a causal inference approach to validate a surrogate by incorporating these longitudinal measurements using a mixed modeling approach for random intercept or random slope models. The value of the surrogate is based on the relationship between the surrogate S and the random effects for T. Based on

these models, we define quantities for surrogacy validation that may vary across the study period using principal surrogacy criteria. We also consider how to define the quantities for surrogacy validation such that they may depend on time. We utilize a surrogate-dependent treatment efficacy curve that allows us to validate the surrogate at different time points, or it is possible to integrate over multiple time points for an overall measure of surrogate validity. Special cases we consider are when T is measured prior to randomization and the delayed-treatment trial design. While in the standard trial design there are non-identified correlation parameters in the complete data likelihood, the potential for crossover treatment arms or use of the pre-treatment measurement allows us to estimate these correlation parameters that arise between treatment arms.

In Chapter IV, we extend these ideas to the surrogate validation framework where both the surrogate marker and the main outcome are time-to-event. Our motivating data source is a localized prostate cancer clinical trial where the two treatments being compared are post-prostatectomy radiation therapy with or without antiandrogen therapy. The trial features men with recurrently elevated prostate-specific antigen (PSA) prostate cancer. The two survival endpoints in the trial are time to distant metastasis and time to death from any cause. New considerations in this setting include the censored nature of these data if patients are lost to follow up or do not experience the events during the study period. Further, it is possible that a patient will experience the terminal outcome without the surrogate endpoint being observed. These considerations complicate the methods proposed within the principal stratification framework so far.

We develop a method that incorporates these issues that arise with multiple survival endpoints. This work considers novel ways to define the parameters measuring the association between outcomes and relevant principal strata. We model conditional hazards while maintaining a valid causal interpretation by viewing this through the lens of a causal multi-state model. In particular, we propose illness death models to accommodate the censored and semi-competing risk structure of

survival data. The proposed causal version of these models involves estimable and counterfactual frailty terms. We propose fixing non-identified parameters using sensitivity analysis. Via these multi-state models, we characterize what a valid surrogate would look like using a causal effect predictiveness plot. We evaluate the estimation properties of a Bayesian method using Markov Chain Monte Carlo using the observed data likelihood and assess the sensitivity of our model assumptions before a concluding discussion.

CHAPTER II

Incorporating Baseline Covariates to Validate Surrogate Endpoints with a Constant Biomarker Under Control Arm

2.1 Introduction

Although randomized clinical trials are largely considered the gold standard to evaluate treatment efficacy, methods that lower trial cost and shorten the length of the study are often sought after in the medical field. In general, surrogate endpoints S are biologically plausible intermediate outcomes that are strongly related to the true outcome of interest T that could act as a substitute for the clinical outcome. In a trial, these endpoints may be measured earlier or more effectively to quickly disperse treatments to patients. Popular examples of potential surrogate endpoints include CD4 blood counts for HIV mortality and immune responses for vaccine efficacy. It is crucial to collect data and validate such an endpoint before using in a large-scale trial. In this paper, we will refer to a surrogate as any intermediate endpoint that occurs between the treatment and measurement of T. Prior to validation, during which the value of potential surrogate is properly assessed, we will use this terminology to mean the surrogate is still a candidate surrogate, whereas we will label it either as a valid or invalid surrogate after such validation procedures.

Prentice's landmark 1989 paper proposed criteria in a single trial setting to evaluate statistical surrogates: that the surrogate S both be related to the outcome and that it captures the effect of the treatment Z on T (Prentice, 1989). Other criteria, such as proportion explained and other causal metrics, have since been proposed, as it has been shown simple criteria may not ensure a seemingly useful surrogate will predict a beneficial treatment effect (VanderWeele, 2013; Freedman et al. 1992). $S(z)$ and $T(z)$ refer to the endpoint values had the treatment, possibly counter-factually, been assigned to level z. Since S is measured after treatment assignment, conditioning on the surrogate distorts the causal pathway and interpretation of the treatment effect in a regression model. Using Rubin's potential outcome causal framework (Rubin, 1974; Little and Rubin, 2000), principal surrogacy proposed the solution of using both potential intermediate outcomes by considering the surrogate values under each treatment as pre-treatment variables (Frangakis and Rubin, 2002). Since both surrogate outcomes are hypothetically determined prior to randomization, surrogates can be evaluated based on principal causal effects. In the case of categorical variables, these are defined on the distribution of $T(1), T(0)$ conditional on principal strata with respect to the pair of posttreatment variables $S(1), S(0)$. Gilbert and Hudgens (forward as GH) defined principal surrogate endpoints based on suggested risk functions of (s_1, s_0) (2008). They elaborated on causal necessity, as proposed by Frangakis and Rubin, by terming average causal necessity and average causal sufficiency. These require that there be no average effect of the treatment on the true outcome in the strata where there is no average effect on the surrogate, and similarly that there exists an average treatment effect on the true outcome in the strata where is there an average effect on the surrogate. Further, they define the Causal Effect Predictiveness (CEP) curve as a visualization of both requirements across the value of the surrogate endpoints (s_1, s_0). The CEP surface is defined based on a chosen contrast function of these risk quantities. In our setting of continuous S and T, we use extensions of the strata formulation using the corresponding quantities $T(1) - T(0)$ and

$S(1) - S(0)$ that are consistent with the previously proposed risk functions.

Our work is motivated by an ongoing study of a muscular dystrophy treatment (Mendell et al., 2020). In this trial of a new gene transfer therapy, patients received a micro-dystrophin transgene to produce the micro-dystrophin protein. The potential surrogate S is micro-dystrophin expression as measured by western blot methods, and the primary outcome of interest T is the North Star Ambulatory Assessment (NSAA) functional score on a continuous scale. S is measured at only one time point, while T is measured before randomization as well as after the gene transfer therapy. Since muscular growth and deterioration due to the disease have major impact on physical movement during childhood, both baseline ambulatory ability and age are important to take into consideration. In this setting, patients do not produce significant amounts of micro-dystrophin protein at baseline, so it can be assumed that the value of the surrogate under placebo is approximately equal to 0. This scenario where $S(0)$ is fixed to 0 also commonly arises in vaccine efficacy studies. Since those in the placebo group necessarily have no immune response without the vaccine antigens, GH refer to this simplified setting as the constant biomarker case.

Quantities related to vaccine efficacy were developed in the HIV and pertussis settings (Halloran, Préziosi, and Chu, 2003; Préziosi and Halloran, 2003; Hudgens and Halloran, 2006), and GH were among the first to formalize the surrogate validation methodology in a HIV vaccine efficacy trial. Still, a major challenge of characterizing these causal effect summaries is dealing with non-identified parameters arising from use of potential outcomes, so Follmann suggested the closeout placebo vaccination design to avoid the unobserved outcomes (2006). GH focused on modeling assumptions to identify the causal quantities and generalized previous work of the baseline immunogenicity predictor (BIP) W to estimate the missing $S(z)$ value. Related work has proposed augmented trial design ideas such as the baseline surrogate measure and the cross-over design, respectively, and other authors have imposed conditional independence assumptions of the outcomes

(Gabriel et al. 2014). Subsequently, several authors have addressed particular models and designs for vaccines and immune correlates of protection (see Gilbert, Qin, and Self, 2008; Wolfson and Gilbert, 2010; Huang, Gilbert, and Wolfson, 2013; Gabriel and Gilbert, 2014; Zhuang, Huang, and Gilbert, 2019; Gilbert and Huang, 2016). In work using principal stratification for a different analytic goal of calculating the average causal effect in a vaccine efficacy study, Shepherd et al. incorporated baseline covariates with the stated purpose of better understanding the mechanism by which the vaccine works (2006).

Under the general Bayesian paradigm, with specific assumptions about parameter values, both Zigler and Belin (2012) and Conlon, Taylor, and Elliott (2014a) proposed to consider the full joint distribution of potential outcomes to create the CEP curve using imputation strategies applicable to settings beyond the constant biomarker case. Work in the frequentist setting by Alonso, Van der Elst, and Meyvisch utilized potential outcomes and the information-theoretic framework to propose a surrogate predictive function with a two-step procedure for dealing with non-identifiability (2017). Making the constant biomarker assumption results in fewer missing potential outcomes to impute, which allows us to focus on the sensitivity of modeling assumptions. In this work, we build upon previous models of the joint distribution of potential outcomes $S(0),\ S(1),\ T(0),\ T(1)$ by exploring trial and modeling considerations when controlling only three of these endpoints as applicable to our motivating clinical trial (Conlon et al. 2014a; Conlon et al., 2017b).

We propose techniques and design considerations with the goal of achieving gains in estimation efficiency. Since the surrogate and true outcome values are only observed for the assigned treatment, we consider the counterfactual outcomes as missing data and implement an imputation strategy for estimation. We compare this algorithm to instead using only observed data and prior distributions for nonidentified correlation parameters. Within our particular goal of surrogate validation, the novelty of this work is the incorporation of baseline covariates with two objec-

tives in mind: first, conditioning on baseline covariates may improve the plausibility of conditional independence assumptions, and second, it allows us to make inference about whether there are subgroups of the population for whom the quality of the surrogate varies. For the latter, we propose to stratify the previously marginal estimands for validation by conditioning on patient characteristics. In our application, we focus on one such example of a baseline covariate, namely that of the true outcome measured pre-treatment, which is similar to the BSM proposed by other authors. This particular measurement allows for multiple definitions of the true outcome of interest and other trial design decisions. We recognize that by viewing the baseline as a noisy estimate of $T(0)$ similar to a measurement error problem, it could provide improved identifiability or yield more informative prior distributions for nonidentified parameters.

In Section 2.2, we propose the model and incorporation of baseline covariates in the surrogate setting. We define the conditional surrogacy validation metrics and suggest potential reasons to use the covariates such as to make conditional independence assumptions and raise consideration for how to define the trial endpoint. In Section 2.3, we describe the proposed Bayesian estimation methods using either an imputation scheme or observed data algorithm. Simulation studies are shown in Section 2.4, and our data example is explored in Section 2.5 before a concluding discussion in Section 2.6.

2.2 The Model

Using the causal association framework, we first consider the joint distribution of three continuous potential outcomes under a binary treatment Z. Since $S(0) = 0$, we assume a multivariate normal

distribution of the counterfactual surrogate and true outcomes for each subject:

$$\begin{pmatrix} S(1) \\ T(0) \\ T(1) \end{pmatrix} \sim MVN \left(\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}, \begin{pmatrix} \sigma^2_{S1} & \rho_{10}\sigma_{S1}\sigma_{T0} & \rho_{11}\sigma_{S1}\sigma_{T1} \\ & \sigma^2_{T0} & \rho_T\sigma_{T0}\sigma_{T1} \\ & & \sigma^2_{T1} \end{pmatrix} \right) \tag{2.1}$$

2.2.1 Assumptions

In this setting, we focus on clinical trial scenarios where we can assume $S(0) = 0$ (the methods can be extended to more general settings). The causal inference assumptions we make are the Stable Unit Treatment Values Assumption and ignorable treatment assignment, meaning the potential outcomes for any individual do not vary with the treatments assigned to other individuals, and treatment assignment is independent of potential outcomes conditional on all covariates ($P(Z = 1|T(0), T(1), X) = P(Z = 1|X)$), respectively. Since we do not observe combinations of joint outcomes $\{T(0), S(1)\}$ or $\{T(0), T(1)\}$, the correlation parameters ρ_{10}, ρ_T are not identified. We will consider various approaches to obtain identifiability through the use of proper priors, conditional independence assumptions, and/or fixing unknown parameters via sensitivity analyses, and consider how our models may be adapted if additional baseline data is available. Then we use the specified joint model to impute the missing counterfactual values. This potential outcomes approach captures the causal associations for validation.

The multivariate normality assumption provides many convenient results by being analytically tractable and allowing for closed form quantities for surrogate validation which we describe below. As this may not hold in practice, we later verify the sensitivity of this distributional assumption and assess the robustness of the results in the presence of model misspecification. Other work has incorporated copula models for non-normal data (Taylor et al. 2015).

2.2.2 Baseline Covariates

Some estimates for surrogacy quality have wide confidence bands that make definitive recommendations difficult. While most causal metrics are reported marginally, it may be beneficial to use baseline covariates X in the analysis. One interest is to assess effect modification: that is, if there exist subgroups of patients to determine for whom the surrogate will work particularly well for the true outcome, such as for males or those who are young. This has the potential to reduce the risk of observing the surrogate fail in a certain patient population after approval for use in subsequent trials. It is also possible that covariates would help predict membership of principal strata. X may explain dependence or confounding between S and T that may occur in finite samples, even after trial randomization. Statistical benefits may be seen in the estimation accuracy as well via more accurate imputation of the missing counterfactual values to both reduce bias and gain efficiency. Finally, conditioning on X might be expected to reduce correlations amongst the potential outcomes, allowing us to make stronger conditional independence statements after conditioning on baseline covariates. Note that these latter examples' effects can be reported as either conditional on X, or integrated over the empirical distributions of covariates to provide marginal estimates.

The conditional model can be written with effects of X in the mean structure (therefore the parameters in this covariance structure differ and θ represents the conditional correlations)

$$\left(\begin{array}{c|c} S(1) & \\ T(0) & X \\ T(1) & \end{array}\right) \sim N\left(\left(\begin{array}{c} \omega_1+\omega_2 X \\ \omega_3+\omega_4 X \\ \omega_5+\omega_6 X \end{array}\right), \left(\begin{array}{ccc} \epsilon_{S1}^2 & \theta_{10}\epsilon_{S1}\epsilon_{T0} & \theta_{11}\epsilon_{S1}\epsilon_{T1} \\ & \epsilon_{T0}^2 & \theta_T\epsilon_{T0}\epsilon_{T1} \\ & & \epsilon_{T1}^2 \end{array}\right)\right) \tag{2.2}$$

We note that this model still has two nonidentified parameters, θ_{10} and θ_T. In the above model X is a scalar, but it could be generalized to a vector. Furthermore, one of the X components might

be known to be highly related to either S or T (such as a pre-treatment measurement). In these specific cases that provide additional information, it may be feasible to make further assumptions about the model structure.

2.2.3 Surrogacy Validation

The validation causal quantities derived from conditioning on strata of the surrogate can be written as a function of the model parameters. These quantities can be viewed graphically in the causal effect predictiveness (CEP) surface as a line with intercept and slope based on causal effects as the difference in surrogate potential outcomes $S(1) - S(0) = s$ on the x-axis and difference in the expected, conditional true outcomes $E(T(1) - T(0)|S(1) - S(0) = s)$ on the y-axis. In the case of Gaussian distributions, as in equation 2.1, this assumption means $E(T(1) - T(0)|S(1) - S(0) = s)$ is linear in s and has the form $= \gamma_0 + \gamma_1 s$. By displaying expected change in potential outcomes, conditional on the actual surrogate change s, the plots demonstrate if the surrogate is valid, meaning small (large) causal effects on a surrogate are associated with small (large) causal effects on the outcome. When the distribution of outcomes is multivariate normal, average causal necessity and average causal sufficiency are fulfilled if $\gamma_0 = 0$, the expected change in true outcome when there is no change in the surrogate outcome at the origin, and $\gamma_1 \neq 0$, the expected change in true outcome when there is a nonzero change in the surrogate outcome. Under the multivariate normal distribution in equation 2.1, these values from the conditional expectation can be written as

$$\gamma_0 = (\mu_{T1} - \mu_{T0}) - \gamma_1 \mu_{S1} = (\delta_3 - \delta_2) - \gamma_1 \delta_1 \qquad \gamma_1 = \frac{\rho_{11}\sigma_{T1} - \rho_{10}\sigma_{T0}}{\sigma_{S1}}$$

or when incorporating baseline covariates, $\gamma_0 = (\mu_{T1|X} - \mu_{T0|X}) - \gamma_1 \mu_{S1|X}$ where $\gamma_1 = \frac{\theta_{11}\epsilon_{T1} - \theta_{10}\epsilon_{T0}}{\epsilon_{S1}}$. Our goal is to estimate all parameters in the distribution so we can calculate γ_0

and γ_1 and determine if S is a valid surrogate for T. We can understand γ_1 by rewriting the quantity as $\frac{\omega_{11}-\omega_{10}}{\sigma_{S1}}$ and can consider when $\omega_{11} > \omega_{10}$ that the slope will be positive, and when the ratio ω_{11}/ω_{10} is larger, the magnitude of γ_1 is larger. Another way to look at this term is as $\frac{Cov(T(1),S(1))-Cov(T(0),S(1))}{Var(S(1))}$, where the sign and magnitude of γ_1 is determined by the covariances between $S(1)$ and the true outcomes.

For scenarios where we believe the surrogate works particularly poorly for certain patient groups, we would be interested in a stratified CEP curve. In other settings, we may simply incorporate covariates as an intermediate step to benefit from the possibility of gains in efficiency by making stronger assumptions while remaining interested in the marginal CEP curve. To denote the difference, let $\gamma_{0,C}$ and $\gamma_{1,C}$ correspond to a model fit using X, compared to the marginalized estimates that is accomplished by empirically averaging over the distribution of X. Later, we will further differentiate these respective conditional C and marginal M models based on how we define the outcome. From equation 2.1, $T(1) - T(0)|S(1), x$ has a normal distribution, so the expected value will be written as $\gamma_{0,C} + \gamma_{1,C}s$. In this expression $\gamma_{0,C}$ can depend on the covariate X, but by the assumptions of the model $\gamma_{1,C}$ (i.e. the covariance) does not depend on X. From these values, the marginalization is written

$$\int_x E(T(1)-T(0)|S(1), X = x)f(X|S(1) = s)dx = \int_x \frac{E(T(1) - T(0)|S(1), x)f(S(1)|x)f(X)}{f(S(1))}dx \tag{2.3}$$

using Bayes rule (see Appendix A). Once we obtain these quantities, we plot the marginal effect over values of s and summarize it by fitting a linear model to estimate $\gamma_{0,M}$ and $\gamma_{1,M}$.

2.2.4 Subgroups in Surrogacy and Treatment Effects

Based on our definition of γ_0, our concept of surrogacy subgroups is analogous to heterogeneous treatment effects existing. Since the quantity $\gamma_{0,C}$ depends on x, a surrogacy subgroup effect will occur only when there is an interaction with the treatment effect on either the surrogate or the true outcome. In contrast, we are not assuming the absence of unobserved heterogeneity in the sense of a sharp null existing (as discussed in Ding, Feller, and Miratrix, 2014).

2.2.5 Conditional Independence

Conditional independence assumptions are frequently made in causal inference and for surrogate endpoint validation in particular (see Conlon et al., 2014a for examples of assumptions made; Parast, Cai, and Tian, 2016; Parast, McDermott, and Tian, 2016; Gilbert et al., 2008, and others). Briefly, when $S(0)$ is not restricted to the value of 0, common assumptions among the endpoints include strong statements about the outcomes such as 1) $T(0) \perp T(1)|S(0), S(1)$; 2) $S(0) \perp T(1)|S(1)$; 3) $S(1) \perp T(0)|S(0)$; and intuitively weaker 4) $S(0) \perp T(1)|S(1), T(0)$ and 5) $S(1) \perp T(0)|S(0), T(1)$.

In our setting, the plausible assumption that is still feasible is a collapsed version of 5: $S(1) \perp T(0)|T(1)$. Determining if a conditional independence assumption is plausible is context-dependent, and we note that this is a potentially weaker assumption than others that have been made in related work such as assuming $T(0)$ and $T(1)$ are conditionally independent (Daniels et al. 2012) Examining the assumption, it is not unreasonable to believe that given the true outcome under $Z = 1$, the surrogate for $Z = 1$ is independent of the true outcome under the opposing treatment. Since we are considering continuous S and T outcomes, we anticipate that $T(1)$ will have sufficient variability and is therefore likely to capture the underlying mechanism of functional

capability in our motivating data example. Making this assumption implies the relationship among equation 2.2 parameters that $\theta_T = \theta_{10}/\theta_{11}$. In the multivariate normal setting, this combination of correlation parameters on the joint normal scale is derived by setting a conditional covariance term to 0 (described in Appendix A), which corresponds to a zero element in the inverse of the correlation matrix. After adjusting for baseline covariates X, conditional independence of the outcomes may be more likely to hold, which in turn reduces the number of parameters to estimate by one (a restatement of equation 2.2 in which θ_{10} has been replaced by $\theta_{11} \times \theta_T$):

$$\left(\begin{array}{c|c} S(1) & \\ T(0) & X \\ T(1) & \end{array}\right) \sim N\left(\left(\begin{array}{c} \omega_1 + \omega_2 X \\ \omega_3 + \omega_4 X \\ \omega_5 + \omega_6 X \end{array}\right), \left(\begin{array}{ccc} \epsilon_{S1}^2 & \theta_{11}\theta_T\epsilon_{S1}\epsilon_{T0} & \theta_{11}\epsilon_{S1}\epsilon_{T1} \\ & \epsilon_{T0}^2 & \theta_T\epsilon_{T0}\epsilon_{T1} \\ & & \epsilon_{T1}^2 \end{array}\right)\right) \tag{2.4}$$

This potentially increases efficiency and helps with identifiability since there is only one nonidentified parameter (θ_T).

2.2.6 Design Considerations and Defining the Endpoint

In our motivating clinical trial, one of our baseline covariates is actually a pre-treatment measurement of the outcome T. Because of this, we could choose to define the true outcomes several ways with the possible benefit of maximizing efficiency. For example, we could define the outcomes $T(0), T(1)$ as the original values of the endpoints and use X for subgroup analysis. Alternatively, we could analyze the outcomes as the change from baseline measurement X to later measurements, denoted as $T^D(0)$ and $T^D(1)$. We would like to know if it is advantageous to define a trial outcome in terms of change from baseline compared to using the baseline value as a covariate. To be explicit about the quantities we are estimating, we can outline the relevant joint and conditional

distributions based on the choice of endpoint definition and involvement of X. All four methods we consider are based on the joint 4x4 distribution from equation 2.1 extended to include X. In the special setting where our covariate is a pre-measurement of $T(0)$, we briefly consider other strong assumptions that could be made. For example, in the strongest case we could plug in X in for $T(0)$ and gain identifiability. Alternatively, we could view X and $T(0)$ as repeated measures and assume certain parameters are equal, such as identifiable means and variances or nonidentified correlations. We consider how these relate to conditional independence assumptions proposed by other authors in future work.

In the setting we are considering, there is a pre-planned analysis of the final clinical trial data, and defining the endpoint is necessary to test for a treatment effect. Like we have suggested for assessing surrogacy, this step involves fitting marginal or conditional models with the original outcome or difference from baseline endpoint. For efficiency or to incorporate treatment effect subgroups, we may choose to condition on baseline covariates X. We enumerate the potential analysis models for the treatment effect based on the observed outcome T and the corresponding surrogate validation models for the potential outcomes $T(0)$ and $T(1)$ conditional on $S(1) = s$. These are differentiated by the endpoint:

1. Original outcome: $T(0), T(1)$ Treatment effect model $T_i = \beta_0 + \beta_1 Z_i + \epsilon_i$

   Surrogate validation metric $E(T(1) - T(0)|s) = \gamma_{0,M} + \gamma_{1,M} s$

2. Also condition on X: $T(0), T(1)|X$ Treatment effect model $T_i = \beta_2 + \beta_3 Z_i + \beta_4 X_i + \epsilon_i$

   Surrogate validation metric $E(T(1) - T(0)|X, s) = \gamma_{0,C} + \gamma_{1,C} s$

3. Difference from baseline: $T^D(0), T^D(1)$ Treatment effect model $T_i^D = \beta_5 + \beta_6 Z_i + \epsilon_i$

   Surrogate validation metric $E(T^D(1) - T^D(0)|s) = \gamma_{0,M} + \gamma_{1,M} s$

4. Also condition on X: $T^D(0), T^D(1)|X$ Treatment effect model $T_i^D = \beta_7 + \beta_8 Z_i + \beta_9 X_i + \epsilon_i$

   Surrogate validation metric $E(T^D(1) - T^D(0)|X, s) = \gamma_{0,C} + \gamma_{1,C} s$

Parameters $\beta_1, \beta_3, \beta_6$, and β_8 estimate the respective treatment effects. We expect the first and third methods to produce estimates of the same population, marginal treatment effect regardless of subtracting off the baseline measurement, though finite sample equality is unlikely to hold for even a randomized trial. These same considerations extend to the surrogate validation framework, where some of these methods will estimate the same marginal validation estimates. We will later consider a further reason to thoughtfully define the endpoint, which is to determine which scale it is reasonable to assume conditional independence.

2.3 Bayesian Methods

2.3.1 Imputation-Estimation Algorithms

While we observe only $S(0), T(0)$ for n_0 subjects and $S(1), T(1)$ for n_1 subjects, our validation quantities involve correlations that can only be calculated from counterfactual outcomes. In order to simultaneously estimate the model parameters, address nonidentified terms, and to appropriately propagate the uncertainty of imputing missing outcomes, we use a Bayesian method for estimation. There are three types of variables that will be iteratively drawn in the MCMC algorithm: the correlation parameters in R, the model mean and variance parameters (μ's and σ's), and the missing potential outcomes. The imputation strategy is a full process that iteratively imputes the missing potential outcomes and uses the posterior distribution to draw values of the parameters.

We assume vague, normal priors for the identified mean parameters. Rather than use an Inverse-Wishart prior or another method that would sample the entire matrix at once that lacks needed

flexibility, we implement a separation method on the covariance matrix Σ. This decomposes the matrix $\Sigma = QRQ$ into standard deviation and correlation matrices where R is a correlation matrix with 1's on the diagonal to easily place less informative priors on the identified terms (Barnard et al. 2000). To ensure iterative draws satisfy the positive definite constraint on Σ, the method uses the griddy Gibbs algorithm to draw from the appropriate bounded posterior (see Conlon et al., 2014a). Specifically, we compute the posterior of each parameter over a set of realizable grid points and re-evaluate over a region of high posterior density with more precision (finer grid points) before randomly drawing the value for that iteration. We consider different priors for the correlation parameter and find that it is important to carefully choose on which parameters to place priors when implementing the conditional independence constraint. Since we are in a setting with nonidentified correlation parameters where the data will provide no direct information about the true values of these parameters, we are careful to not impose unreasonable prior distributions as we expect the posterior to mimic the prior. We consider both vague and more informative Uniform and Beta priors on the correlation terms, though the marginal distribution of each correlation under positive definite constraints can become less straightforward. We assume $S(1) \perp T(0)|T(1)$ (or its equivalent based on the exact model fit and incorporation of X), by drawing suitable values of θ_T and θ_{11} using the grid search. Essentially, as demonstrated in equation 2.4, the term θ_{10} is no longer involved in the likelihood, and the product $\theta_T \times \theta_{11}$ takes its place.

2.3.2 Observed Data Algorithm

An alternative way to estimate the identifiable parameters is by using the observed data likelihood and devising an MCMC algorithm to obtain draws from the posterior distribution. This approach avoids imputing the counterfactual values of S and T. Since the data contain no information about θ_T or θ_{10}, we expect the posterior to match the distribution of the chosen prior, provided the priors

are independent. Thus, we propose to draw the nonidentified correlation parameters directly from the prior solely for the purposes of estimating γ_0 and γ_1. Since the value of θ_{11} is estimable from the observed data, we use the posterior distribution for its draw, but it is the only correlation parameter drawn from a conditional distribution. When assuming conditional independence, we replace the term θ_{10} with $\theta_T \times \theta_{11}$ (after θ_{11} is drawn from the posterior and θ_T is drawn directly from Uniform(-1, 1)) in the same way as explained above when calculating values of γ_1 and ensuring the matrix is positive definite. We then fit the regression models on the observed data only to estimate the other parameters.

Here we carry out basic Bayesian estimation of the identified parameters for comparability of uncertainty estimates to isolate the effect of using the imputation scheme and priors for non-identified parameters. We find that we can bypass the full MCMC scheme intended to provide parameter estimates while addressing the nonidentified parameters, and instead we can draw these independently from the prior. This is related to work by Gustafson (2009) that demonstrates any difference between the prior and posterior distribution for non-identified parameters is due to prior dependence in the parameters. Using his transparent reparameterization here, there is no indirect learning of the correlations outside of the positive definite and conditional independence assumption constraints, which are still enforced with this algorithm. In this setting, we expect results from this method to be generally equivalent to imputation while being less computationally expensive. We also note that using an MCMC scheme to estimate the variance and mean parameters may not be necessary at all, and a maximization of the posterior or maximum likelihood method may be used instead.

2.4 Simulation Studies

We explore the impact of using a baseline covariate in terms of efficiency and making conditional independence assumptions. In particular, our simulations are meant to assess how we define the true endpoint based on different relationships with the baseline covariate. Using simulation studies, we generate data that mimics a randomized trial, meaning for half of the subjects assigned $Z = 0$, we observe only $T(0), X$, and for the other half we observe $S(1), T(1), X$. The six sets of generative parameters for equation 2.2 are shown in Table 2.1. The six settings have different generating parameters that vary the treatment effect and quality of the surrogate endpoint to demonstrate the method's performance over a variety of scenarios. Based on the model in equation 2.2 for observed data of sample size $n = 100$, we generate data from parameter combinations that allow us to assess different settings expected to occur during validation.

The surrogate is valid marginally only for settings A, B. The covariate X is normally distributed in settings A-E and binary in F; settings D, F represent the existence of a subgroup effect. After generating the data, we fit the models described below and vary which conditional independence assumption is made, if any. To perform surrogacy validation, we fit four models of tri-variate normal distributions derived from the distribution of $S(1), T(0), T(1), X$ (see Appendix B for details). The four analysis designs are based on the distribution of the three outcomes and baseline covariates, and the different parameterizations can be easily equated algebraically. For any X (we simulate $X \sim N(\delta_4, \sigma_X^2)$ or $X \sim Bernoulli(0.5)$),

1. $\begin{pmatrix} S(1) \\ T(0) \\ T(1) \end{pmatrix}$

2. $\left(\begin{array}{c|c} S(1) & \\ T(0) & X \\ T(1) & \end{array} \right)$

3. $\left(\begin{array}{c} S(1) \\ T^D(0) \\ T^D(1) \end{array} \right)$

4. $\left(\begin{array}{c|c} S(1) & \\ T^D(0) & X \\ T^D(1) & \end{array} \right)$

We consider that conditional on baseline covariates based on context, we may decide to introduce strong modeling assumptions. For example, when X is a pre-treatment measurement of T (like a measurement-error prone value of $T(0)$), priors may be informed by estimates of the observed correlation of X and $T(1)$, or we may be able to estimate some nonidentified correlations. Currently we use a Beta prior (truncated between -0.4 to 1 with a positive mean equal to 0.23) on the correlation between either $T(0), T(1)$ or $T^D(0), T^D(1)$ and a Uniform(-1, 1) prior on the correlation between $S(1), T(1)$ or $S(1), T^D(1)$ when conditional independence is not assumed. To perform sensitivity analyses, we will both vary the prior distributions on the nonidentified parameters to integrate over the range of plausible values and fix the correlations to see at what boundaries the conclusions change.

These models allow us to contrast the estimation of marginal quantities $\gamma_{0,M}$ and $\gamma_{1,M}$ and also those for subgroups of participants based on baseline covariates X from models 2 and 4. For our purposes, we first compare the marginal estimates for γ_1 and γ_0 which are directly calculated in

settings 1 and 3 and are derived by marginalization for models 2 and 4. The definition of the endpoints (either the difference from baseline D and corresponding $\gamma_{.D}$ or the original value O and $\gamma_{.O}$) is important as $\gamma_{1,D} \neq \gamma_{1,O}$ when there is an effect of X. Further, the validity of conditional independence varies in our data generating mechanism directly by changing the true values of the correlations so that we can assess how the method performs when this is violated. In particular, the data generating mechanism directly violates the conditional independence assumption in setting E when we condition on X and use the original endpoint. The assumption is also violated indirectly in certain other settings when we fit marginal models or take the difference from baseline. For example, when $T(0) \perp S(1)|T(1), X$, it is also true that $T(0)^D \perp S(1)|T(1)^D, X$. However, this corresponding satisfaction of conditional independence does not necessarily hold between $T(0) \perp S(1)|T(1)$ and $T(0)^D \perp S(1)|T(1)^D$ when we do not condition on X. To quantify this discrepancy when this condition is not met, we calculate the deviation from meeting the conditional independence requirement, i.e. how incorrect it is to make this conditional independence assumption (see Appendix C).

2.4.1 Simulation Results

We present the posterior mean point estimates averaged over 1,000 datasets for each setting using the observed data algorithm. To assess variability, we also report the standard deviation of the Bayesian point estimates and the average standard error of the parameter value (the posterior standard deviation) within each replication. We run each MCMC for 3,000 iterations and ensure convergence is reached using traceplots. Using this method, computation is fast enough to run the algorithm in parallel over many replicates in a few hours. Since our goal is to effectively validate surrogate endpoints, we focus on inference for the quantities γ_0 and γ_1. The results show the contrast in estimation accuracy and efficiency when the constraints are enforced during estimation

compared to when no assumptions are imposed during the MCMC procedure.

2.4.2 Marginal Estimates

Below [Figure 2.2] (and in the Appendix C) are results with sample size $n = 100$ using a truncated Beta$(5, 6)$ prior. Bias for rows 1 and 2 as well as 3 and 4 is calculated based on the same true, marginal values. Overall, the identified mean and variance parameters are estimated with little bias. However, the nonidentified correlations are sensitive to the prior distribution. Compared to supplementary tables in Appendix C, for comparability across settings Figure 2.2 has been adjusted by the empirical variability in the validation estimates from complete counterfactual data: we calculated the maximum likelihood estimates under a scenario where we would observe all n counterfactual outcomes and fit a model for $T(1) - T(0)$ conditional on $S(1)$. We adjusted the bias and standard errors relative to the standard deviation of the estimates across 1,000 simulations to provide quantities that can interpreted in proportion to the amount of variability in the data and the estimates.

These results show there can be reduced bias in the estimates when making conditional independence assumptions such as in settings B and C. Notably, the credible intervals for the nonidentified parameters are very conservative for all settings as expected due to the non-identified correlation parameter and its associated relatively weak prior, and the corresponding coverage probabilities are near one. Further, the within simulation average, over-estimation of the standard error decreases when making conditional independence assumptions, though the SD of the estimates is not necessarily smaller when making these assumptions (through less under-identification in the conditional independence model, this reduction in SE improves the agreement between the average SE and SD). This is seen in Figure 2.2 when comparing the average standard errors, as shown in the dotted lines, to the usually smaller standard deviation of the point estimates, as seen in the solid,

shaded lines. The scenarios assuming conditional independence in red have better correspondence than the scenarios without the assumption during model fitting, shown in blue. We see that conditioning on X can also decrease the standard error of the estimates, particularly in the scenario where we do not make conditional independence assumptions in settings B, D, and E for example. In other settings, the difference is not as clear. In setting E, we also see very directly the impact of assuming conditional independence when it does not hold. In the case where we use the original endpoint, the bias is slightly reduced when we do not make the conditional independence assumption.

2.4.3 Sensitivity to Distributional Assumptions

We show results in Appendix E for fitting conditional models when the outcomes are heavy tailed (t-distributed) or skewed (gamma-distributed). When data follows these non-Gaussian distributions, the estimation results do vary from previous simulations as seen in the point estimates and their corresponding variability intervals. While there is some increased bias and variability of the estimates as compared to those with a normal distribution, generally there is some robustness in the estimation such that the conclusions regarding surrogate validity (i.e. if the credible interval for γ_0 covers zero while that for γ_1 does not) seem to be similar in the settings we considered. We further explore the role of distributional assumptions and limitations in the discussion.

2.4.4 Subgroup Analysis

Now we focus on the results of simulation setting E where subgroups exist for the models fit using methods 2 and 4 (conditional on X). Since X is binary in this setting, we report the conditional values of γ_0 when $X = 0$ and $X = 1$. [Table 2.2] Whereas Figure 2.2 shows marginalized

estimates of the quantities, Table 2.3 shows the estimates of γ_0 and γ_1 conditional on the values of X, which allows us to determine if the surrogate is valid for patients with certain baseline characteristics. Looking at the validity of S as a surrogate for subgroups in setting F, we see that the surrogate is valid for $X = 0$ but not for $X = 1$ by examining the estimates of γ for the values of X separately. If X were gender, for example, this would indicate that the surrogate is valid only for males and not females. Our validation criteria state S is a valid surrogate when the credible interval for γ_0 covers 0 and the interval for γ_1 does not. Table 2.3 shows the proportion of simulations for which this is the case. The criteria that S is a valid surrogate for the group $X = 0$ is only detected for the model where we make a conditional independence assumption. When we do not, the credible interval of γ_1 overlaps with 0 in almost all simulation replications, signaling that we are not able to declare the surrogate as valid for any individuals.

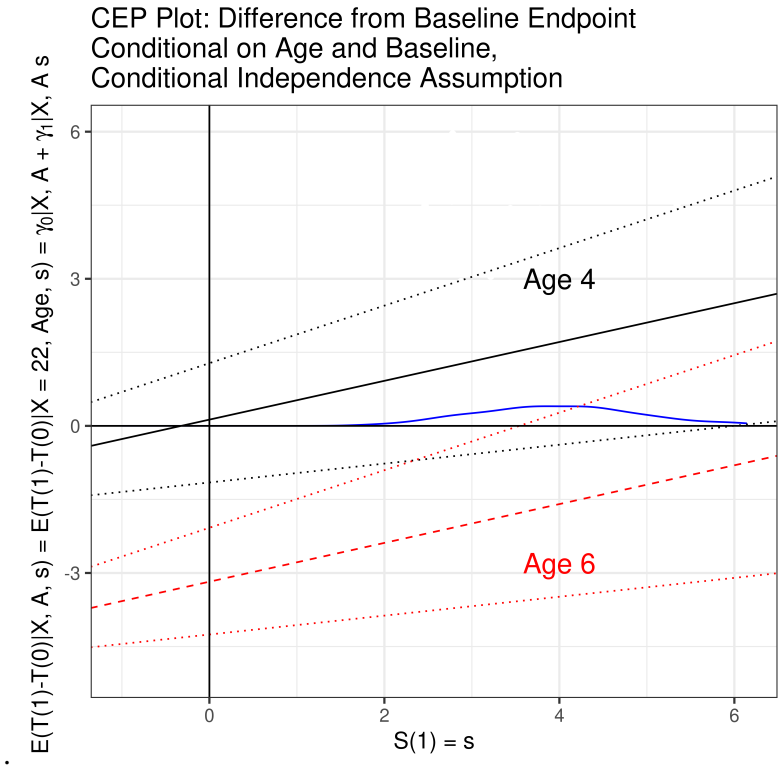
2.5 Duchenne Muscular Dystrophy Data Example

We generate "observed" data for three normally-distributed outcomes aimed to mimic an ongoing clinical trial and the natural progression and deterioration of ambulatory function for Duchenne muscular dystrophy patients (Muntoni et al. 2019). Since the trial has not been unblinded, the parameter values were chosen to match the preliminary data when available, and for other parameters the values were chosen in a subjective way to be what the authors considered as reasonable. Based on the literature and company analysis standards, to assess surrogacy, we condition on age at baseline, A, and the baseline NSAA score measurement, X.

2.5.1 Data Example Results

First we show the estimated treatment effects that would be calculated in a clinical trial measuring the efficacy of the treatment based on the true, observed endpoints T. [Table 2.3] It is clear that the standard error estimates for γ_0 and γ_1 are markedly smaller, and credible intervals more narrow, when we make the conditional independence assumptions. We can see that marginally, the surrogate of micro-dystrophin is not a valid surrogate for improvement of the NSAA score across the entire study sample. However, there are strong effects of both age and baseline measurement on the outcome. We can identify a region of the covariate space based on age and baseline measurement where the surrogate is valid for a subgroup of patients. We create CEP curves for these data to show what the surface looks like when we stratify based on covariates. [Figure 2.1]

Figure 2.1: The CEP plot shows the conditional functions of γ_0 and γ_1 across possible values of $S(1)$. The empirical distribution of $S(1)$ is shown in the blue density curve.



The CEP plot is conditional on baseline NSAA and age, and we demonstrate that the surrogate is valid for those at four years of age. Since age is modeled with linear and quadratic effects, the surrogate will be invalid for those six and older as the estimated CEP curve moves farther from 0. This demonstrates that due to the combination of natural growth and degeneration due to disease over time, the surrogate would only be valid within a certain younger patient population.

We also explore the consequences of different prior distributions for the non-identified parameter θ_T and compare this to fixing θ_T at some value, shown in supplementary figures in Appendix D. We see that the results for γ_0 and γ_1 are somewhat sensitive to the choice, but the conclusion that the surrogate is valid holds for values of θ_T which we believe to be reasonable. The results are the same for both the imputation and observed data algorithms.

2.6 Discussion

In this work, we have focused on incorporating baseline covariates into the validation process for surrogate endpoints. Our motivation for including such covariates to assess surrogacy is to potentially increase efficiency through the use of modeling assumptions and to allow for the possibility of heterogeneity in the utility of the surrogate endpoint across patients. Considering the harmful implications of incorrectly validating a surrogate endpoint, it may be worthwhile to consider the CEP as a function of X and identify potential subgroups of patients for which a surrogate is appropriate.

While we have identified scenarios where there are gains by incorporating baseline covariates, there are some situations where efficiency improvements are limited. Further, when implementing the proposed model assumptions, it is important to assess the plausibility of conditional independence assumptions even in this context of adjusting for baseline covariates which we believe

makes the assumptions more likely. Introducing these conditional independence constraints aids in improved estimation properties, but they should not be implemented without proper, context-dependent reasoning. Further, there are many ways to implement the constraints, although these simulations suggest certain strategies (using observed data only) may reduce the burden of imputing potential outcomes but still require reasonably well-specified prior distributions. There are many ways in which these methods and simulations will be extended, particularly to verify its robustness. Work is ongoing to incorporate baseline covariates and conditional independence assumptions into this framework while fully utilizing the longitudinal data. In the longitudinal study design setting, the potential for crossover treatment arms or use of the pretreatment measurement allows for more direct incorporation of potentially identified correlation parameters that arise between treatment arms.

A limitation of the proposed method is that it relies on a well-specified model to validate surrogate endpoints. Using the CEP curve with the normality assumption results in a linear form for the conditional expectation of interest, though the CEP concept itself does not rely on such a distributional assumption, and more flexible copula or other modeling could be implemented instead (Taylor et al. 2015; Kim et al. 2017; Ma et al., 2011). Making the normality assumption is a helpful first step to develop surrogacy metrics in closed-form. Semi-parametric methods could be employed to assess surrogacy under more flexible models. Here we have briefly explored the method's performance under model misspecification where the outcomes are not multivariate normal, and we have found some sensitivity to this in our results. In general, the method does still demonstrate reasonable performance as far as suggesting the appropriate surrogacy conclusions in the considered scenarios. Other extensions that do not rely on the multivariate normality assumption include methods for time-to-event data using different assumptions to estimate subject-specific $\Delta S_i = S(1)_i - S(0)_i$ and $\Delta T_i = T(1)_i - T(0)_i$ and empirically fitting CEP curves to these es-

timates. Finally, principal stratification and CEP curves are one of many strategies to determine the validity of surrogate endpoints. Future investigation into potential correspondence between different assumptions of models as done in Conlon et al. (2014a) for more metrics such as proportion explained may be informative in comparing the methods and their sensitivities to modeling assumptions.

2.7 Publication

The content of this chapter has been published in *Statistics in Medicine* at DOI: 10.1002/sim.9201

2.8 Figures

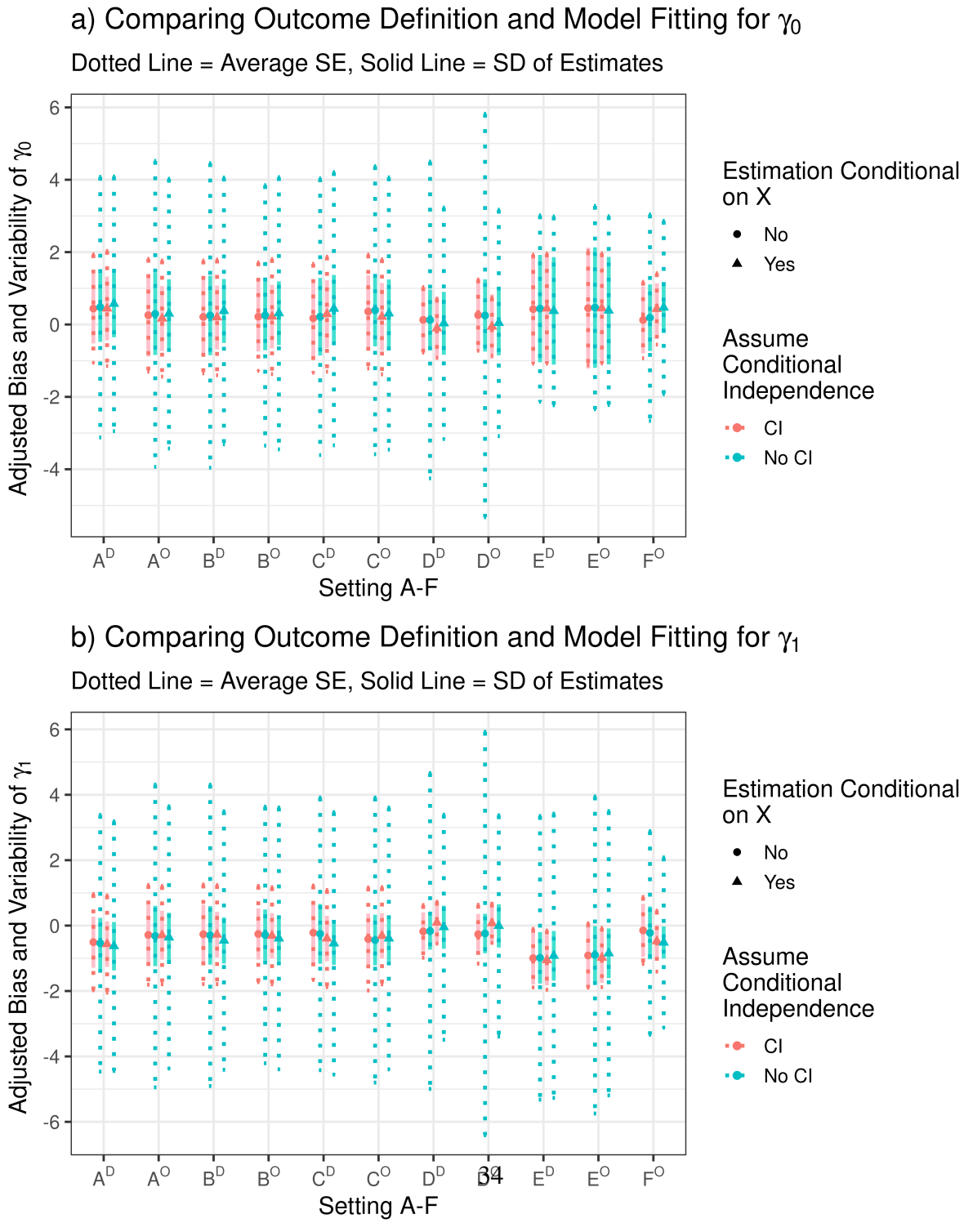
	A	B	C	D	E	F
	Valid S	Valid S: No X Effect	Invalid S	Valid S for SG	Invalid CI	Valid S for SG
σ_X	0.5	0.5	0.5	0.5	0.5	NA
δ_4	1	1	1	1	1	NA
ω_1	2	2	2	2	0.5	2
ω_2	0	0	0	0	0	0
ω_3	3	3	3	3	3	3
ω_4	1	0	1	3	1	-0.75
ω_5	4.1	4.1	4.1	4.1	4.1	4.1
ω_6	1	0	1	1	1	2
$\epsilon_{S1} = \epsilon_{T0} = \epsilon_{T1}$	1	1	1	1	1	1
θ_{10}	0.15	0.15	0.15	0.08	-0.05	0.15
θ_{11}	0.7	0.7	0.7	0.3	0.25	0.7
θ_T	0.21	0.21	0.21	0.26	0.21	0.21
$\gamma_{0,O}$	0	0	-1.00	-1.35	0.95	1.31
$\gamma_{1,O}$	0.55	0.55	0.55	0.22	0.30	0.58
$\gamma_{0,D}$	-0.06	0	-1.02	-1.33	0.95	NA
$\gamma_{1,D}$	0.58	0.55	0.56	0.22	0.31	NA

Table 2.1: Generative parameter values for the six scenarios to compare definition of the endpoint and using baseline covariates. SG stands for subgroup based on X. The meaning of these parameters is shown in equation 2.2. Note 'difference from baseline' (subscript D) is not defined when X is binary and T is continuous (F).

Setting	Fit Conditional Independence Assumption	$\gamma_{0,C}$ True $X=0$	Est	SE	SD	Covers 0	$\gamma_{0,C}$ True $X=1$	Est	SE	SD	Covers 0	$\gamma_{1,C}$ True	Est	SE	SD	Covers 0
2F	$T(0) \perp S(1) \mid T(1), X$	0	0.068	0.465	0.307	0.987	2.75	2.797	0.460	0.319	0.000	0.55	0.514	0.176	0.087	0.003
2F	None	0	0.090	0.984	0.308	1.000	2.75	2.808	0.985	0.319	0.040	0.55	0.503	0.466	0.088	0.978

Table 2.2: Simulation results demonstrating effect of estimating subgroups. Estimates of γ_0 and γ_1 are conditional on the values of X, so values of γ_0 estimates are conditional on $X = 0, 1$. The columns denoting 'covers 0' indicate what proportions of simulations have credible intervals that do contain 0. This helps determine for how many simulations the surrogate would be considered valid. For γ_0, the credible interval covering 0 denotes a valid surrogate, while the interval of γ_1 should not.

Figure 2.2: Simulation results for the two definitions of the endpoint (here categorized by superscript to denote O = Original Endpoint (Settings 1 and 2) compared to D = Difference from baseline (Settings 3 and 4)) and different generating parameter values $(A - F)$. The values shown below are the bias and variability (both the average within-sample standard error across datasets and the standard deviation of the point estimates) of the validation quantities that are adjusted by the variability of the hypothetical surrogate values if all counterfactuals were to be observed (from the full data) using the standard deviation of point estimates from standard regression models for $T(1) - T(0)|S(1)$.



Endpoint Type	Treatment Effect	Conditional Independence Assumption	γ_0 Estimate	γ_0 SE	γ_1 Estimate	γ_1 SE
T^D	0.227	$T^D(0) \perp S(1) \vert T^D(1), X, A$	-2.215	0.408	0.396	0.101
		None	-2.198	1.107	0.391	0.277
T	0.271	$T(0) \perp S(1) \vert T(1), X, A$	-2.218	0.409	0.397	0.102
		None	-2.194	1.109	0.391	0.278

Table 2.3: Simulated muscular dystrophy estimates of treatment effect and marginal surrogacy quantities.

CHAPTER III

Solutions for Surrogacy Validation with Longitudinal Outcomes for a Gene Therapy

3.1 Introduction

Valid surrogate endpoints, which we will refer to as S, can be used as substitutes for a true outcome of interest T to measure the efficacy of treatment Z in a clinical trial. Surrogate endpoints can lower clinical trial costs and shorten study lengths. Unfortunately, using an inadequate endpoint can lead to inaccurate conclusions regarding treatment effects. Statistical criteria for validation have been proposed since Prentice's landmark paper (Prentice, 1989), including causal inference approaches where $S(z)$ and $T(z)$ refer to the endpoint values had the treatment been assigned to treatment level $Z = z$. Principal surrogacy is one causally-valid solution where the treatment Z and counterfactual values of S and T are jointly modeled (Frangakis and Rubin, 2002). In our setting of multivariate normal endpoints, principal causal effects are defined on the distribution of $T(1) - T(0)$ conditional on $S(1) - S(0)$. Building upon these ideas, Gilbert and Hudgens defined the Causal Effect Predictiveness (CEP) curve to verify the quality of the surrogate within this framework (2008).

In our motivating muscular dystrophy trial with a binary therapy Z, with $Z = 1$ denoting the

gene therapy and $Z = 0$ denoting the placebo, T is a measure of the mobility and strength of an individual. The candidate surrogate S is an expression of micro-dystrophin that is measured a few weeks after the therapy is initiated. A characteristic of the disease is a lack of protein due to micro-dystrophin, and the gene therapy is aiming to activate this gene. The outcomes $T(0), T(1)$ are measured longitudinally, and we can assume $S(0) = 0$ since subjects with the disease and without gene therapy will have essentially zero gene expression. Gilbert and Hudgens formalized surrogate validation methods for a vaccine efficacy trial with non-longitudinal outcomes and refer to the setting where placebo participants have no immune response (i.e. $S(0)$ is fixed to 0) as the constant biomarker case. Work by Roberts, Elliott, and Taylor (2021) incorporates baseline covariates in this setting for a trial with cross-sectional measures. The natural history of ambulatory ability due to disease progression has been characterized by Muntoni et al. (2019). Since muscular growth and deterioration from disease have major impact on mobility, effects of time are important to consider when evaluating surrogacy. Further, the trial design includes a cross-over or delayed-start treatment portion, so it may be possible to measure otherwise counterfactual T and S outcomes since placebo subjects receive the treatment mid-trial.

Most existing literature on time-varying effects for surrogate validation describe joint models where either the treatment or surrogate is time-varying or repeatedly measured (Hsu et al., 2015; Agniel and Parast, 2020) without addressing our situation of repeated true outcomes. To our knowledge, there has not been work in this setting where the true outcome is repeatedly measured in the context of surrogacy validation for a single trial. Validation metrics have been proposed for multiple trials with repeated measures using meta-analysis (Alonso et al., 2003; Alonso et al., 2004; Renard et al., 2003). Related work by Gabriel and Gilbert (2014) introduces the time- and surrogate-dependent treatment efficacy curve for time-to-event data. This work makes inference on the time-varying value of S as a surrogate using the estimated causal quantities over a range of

time points. In subsequent research, Gabriel, Sachs, and Gilbert (2015) incorporate time for their proposed standardized total gain metric for censored data and suggest it can be integrated over time.

Similar trials with delayed-start treatment, closeout, or stepped wedge designs have been considered in the literature (Follmann, 2006; Brown and Lilford, 2006). Validation methods for a vaccine closeout trial in Qin et al. (2008) assume time constancy where the closeout measurement is equal to the true value of S plus measurement error. Luedtke and Wu (2020) similarly assume the true outcomes for the crossover individuals are equal to what would have been observed had they received treatment at baseline. We will use crossover to denote the time that an individual moves from the placebo to treated arm. A strength of our proposed mixed model is its allowance for time-dependence, which addresses the concern that $T(1)$ in the delayed-treatment group individuals may not be identical to what would be observed at time of randomization without controlling for time.

In this paper, we build upon work for the joint distribution of normally-distributed potential outcomes by incorporating observed longitudinal outcomes for T and focus on the setting where $S(0) = 0$. We extend established estimands for validation based on the conditional expectation $E(T(1) - T(0)|S(1) - S(0) = s)$ where the desired features of a surrogate are for this quantity to be close to zero when $s = 0$, and far from zero when s is not equal to zero. We explore how this distribution and validation metric can allow for repeated measures. Further, special cases we consider are when T is also measured prior to randomization and a delayed-start of treatment trial design. We model these additional endpoints in our proposed validation framework to demonstrate the potential benefit of these designs.

3.2 The Model

We first consider the joint distribution of three potential outcomes under the causal association framework. Previous work with cross-sectional data assumed a multivariate normal (MVN) distribution of the counterfactual surrogate and true outcomes by parameterizing the model in the following way (Conlon, Taylor, and Elliott, 2014a; Roberts et al., 2021)

$$\begin{pmatrix} S(1) \\ T(0) \\ T(1) \end{pmatrix} \sim MVN \left(\begin{pmatrix} \alpha_1 \\ \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \sigma^2_{S1} & \theta_{10}\sigma_{S1}\sigma_{T0} & \theta_{11}\sigma_{S1}\sigma_{T1} \\ & \sigma^2_{T0} & \theta_T\sigma_{T0}\sigma_{T1} \\ & & \sigma^2_{T1} \end{pmatrix} \right) \tag{3.1}$$

Here the population means are represented by $\{\alpha_1, \beta_0, \beta_1\}$, σ^2 terms denote variances, and θ parameters indicate correlations among the outcomes. In a standard trial design, where the patients are randomly assigned to either $z = 1$ or $z = 0$ arm, both correlations θ_T and θ_{10} are non-identified.

3.2.1 Repeated Measurements in a Standard Trial Design

We extend the above MVN model to the situation where each individual $i = 1, ..., n$ has several observed and counterfactual outcomes $S(1)_i, T(0)_{ij}, T(1)_{ij}$ for $j = 1, ..., m_i$ repeated measures in a randomized trial. In this paper, we derive distributions where m is equal for all individuals and $S(1)_i$ is only measured once, and is therefore a scalar, but this could be relaxed. In the simplest random intercept model, we propose that each individual has one random effect for the vector of $\boldsymbol{T(0)}$ measurements and one for $\boldsymbol{T(1)}$ measurements, denoted $b^{(0)}$ and $b^{(1)}$ respectively. Then, each measurement $T(z)_{ij}$ for subject i and time j is assumed to follow a linear mixed model:

$$T(0)_{ij} = \boldsymbol{\beta}_0 \boldsymbol{X}_{ij} + \boldsymbol{Z}_{ij} b_{.i}^{(0)} + e_{ij0} \qquad T(1)_{ij} = \boldsymbol{\beta}_1 \boldsymbol{X}_{ij} + \boldsymbol{Z}_{ij} b_{.i}^{(1)} + e_{ij1} \qquad e_{ijz} \sim N(0, \sigma_e^2) \quad (3.2)$$

and $S(1)_i = \alpha_1 + \epsilon_i, \epsilon_i \sim N(0, \sigma_{S1}^2)$ with baseline or time-dependent covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$. Later we will discuss the special case when there is also a pre-treatment measurement of T.

To extend the structure in (1), we assume that the joint distribution of $(S(1)_i, \boldsymbol{T(0)}_i, \boldsymbol{T(1)}_i)$ is determined by the distribution of $(S(1)_i, b_i^{(0)}, b_i^{(1)})$ and consider the joint distribution of the surrogate and the random effects for a random intercept model:

$$\begin{pmatrix} S(1)_i \\ b_i^{(0)} \\ b_i^{(1)} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \alpha_1 \\ 0 \\ 0 \end{pmatrix}, \Psi = \begin{pmatrix} \sigma_{S1}^2 & \rho_{10}\sigma_{S1}\sigma_{b0} & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ & \sigma_{b_0}^2 & \rho_T \sigma_{b_0}\sigma_{b_1} \\ & & \sigma_{b_1}^2 \end{pmatrix} \right) \quad (3.3)$$

The corresponding distribution of outcomes conditional on random effects then is written

$$\left(\begin{array}{c|c} S(1)_i & \\ \boldsymbol{T}(0)_i & b_i^{(0)} \\ \boldsymbol{T}(1)_i & b_i^{(1)} \end{array} \right) \sim MVN \left(\begin{pmatrix} \alpha_1 + \delta_1 b_i^{(0)} + \delta_2 b_i^{(1)} \\ \boldsymbol{\beta}_0 \boldsymbol{X}_j + \boldsymbol{Z}_i b_i^{(0)} \\ \boldsymbol{\beta}_1 \boldsymbol{X}_j + \boldsymbol{Z}_i b_i^{(1)} \end{pmatrix}, \begin{pmatrix} \sigma_{S1}^2 - \delta_3 & 0 & 0 \\ & \sigma_e^2 & 0 \\ & & \sigma_e^2 \end{pmatrix} \right)$$

where $\delta_1 = \frac{(\rho_{10} - \rho_{11}\rho_T)\sigma_{S1}}{\sigma_{b0} - \rho_T^2 \sigma_{b0}}$, $\delta_2 = \frac{(\rho_{11} - \rho_{10}\rho_T)\sigma_{S1}}{\sigma_{b1} - \rho_T^2 \sigma_{b1}}$, and $\delta_3 = \frac{-(\rho_{10}^2 + \rho_{11}^2 - 2\rho_{10}\rho_{11}\rho_T)\sigma_{S1}^2}{\rho_T^2 - 1}$.

In our trial, we will model age or time over the course of the trial using random slopes. Consider the random intercept and random slopes models for $T(0)$ and $T(1)$ in equation 3.4 where the random effect vectors and covariance matrices increase in dimensionality.

$$\begin{pmatrix} S(1)_i \\ b_{i\cdot}^{(0)} \\ b_{i\cdot}^{(1)} \end{pmatrix} = \begin{pmatrix} S(1)_i \\ \begin{pmatrix} b_{i0}^{(0)} \\ b_{i1}^{(0)} \end{pmatrix} \\ \begin{pmatrix} b_{i0}^{(1)} \\ b_{i1}^{(1)} \end{pmatrix} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \alpha_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi} \right), \boldsymbol{\Psi} = \begin{pmatrix} \sigma_{S1}^2 & \sigma_{S1}\sigma_{b01}\rho_{S101} & \sigma_{S1}\sigma_{b11}\rho_{S111} & \sigma_{S1}\sigma_{b02}\rho_{S102} & \sigma_{S1}\sigma_{b12}\rho_{S112} \\ & \sigma_{b01}^2 & \sigma_{b01}\sigma_{b11}\rho_{0111} & \sigma_{b01}\sigma_{b02}\rho_{0102} & \sigma_{b01}\sigma_{b12}\rho_{0112} \\ & & \sigma_{b11}^2 & \sigma_{b11}\sigma_{b02}\rho_{1102} & \sigma_{b11}\sigma_{b12}\rho_{1112} \\ & & & \sigma_{b02}^2 & \sigma_{b02}\sigma_{b12}\rho_{0212} \\ & & & & \sigma_{b12}^2 \end{pmatrix} \tag{3.4}$$

We denote a baseline covariate B_i and time from randomization using the vector $\boldsymbol{t}_i$. For each individual, $\boldsymbol{T(0)}_i, \boldsymbol{T(1)}_i$ have random and fixed effects where $\boldsymbol{Z}_i = (1 \quad \boldsymbol{t}_i)$, $\boldsymbol{X}_i = \begin{pmatrix} 1 & B_i \times \mathbf{1} & \boldsymbol{t}_i \end{pmatrix}$, $\boldsymbol{\beta}_0 = (\beta_0^{(0)} \quad \beta_1^{(0)} \quad \beta_2^{(0)})$, $\boldsymbol{\beta}_1 = (\beta_0^{(1)} \quad \beta_1^{(1)} \quad \beta_2^{(1)})$, and each $T(z)_{ij}$ has a distribution that depends on time. Letting each δ be a function of covariance parameters in $\boldsymbol{\Psi}$, we make the assumption that

$$\left(\begin{array}{c|c} S(1)_i & b_{i0}^{(0)}, b_{i0}^{(1)}, \\ \boldsymbol{T(0)}_i & b_{i1}^{(0)}, b_{i1}^{(1)}, \\ \boldsymbol{T(1)}_i & B_i, \boldsymbol{t}_i \end{array} \right) \sim MVN \text{ with mean } \begin{pmatrix} \alpha_1 + \delta_1 b_{i0}^{(0)} + \delta_2 b_{i0}^{(1)} + \delta_3 b_{i1}^{(0)} + \delta_4 b_{i1}^{(1)} \\ \beta_0^{(0)} + \beta_1^{(0)} B_i + \beta_2^{(0)} \boldsymbol{t}_i + b_{i0}^{(0)} + b_{i1}^{(0)} \boldsymbol{t}_i \\ \beta_0^{(1)} + \beta_1^{(1)} B_i + \beta_2^{(1)} \boldsymbol{t}_i + b_{i0}^{(1)} + b_{i1}^{(1)} \boldsymbol{t}_i \end{pmatrix}. \tag{3.5}$$

3.2.2 Pre-treatment Observations of T

We consider that the covariates may have a special form and explore the use of a pre-treatment observation of T, denoted by T_{BL_i}. When this additional data is available, we can either include this as a baseline covariate or treat it as an outcome measure of $T(0)$ at $t = 0$ in our mixed model. This extension may help with identifiability, because both T_{BL} and $T(1)$ are measured in the same person, giving information about the previously non-identified parameters $\rho_{0102}, \rho_{0112}, \rho_{S101}$. If we chose to model T_{BL} as a covariate, after fitting the conditional model we will have to integrate over T_{BL} to obtain marginal quantities of our validation metrics. These ideas are explained in more detail in Section 3.3 and in Appendix H.

3.2.3 Delayed-Start of Treatment Design

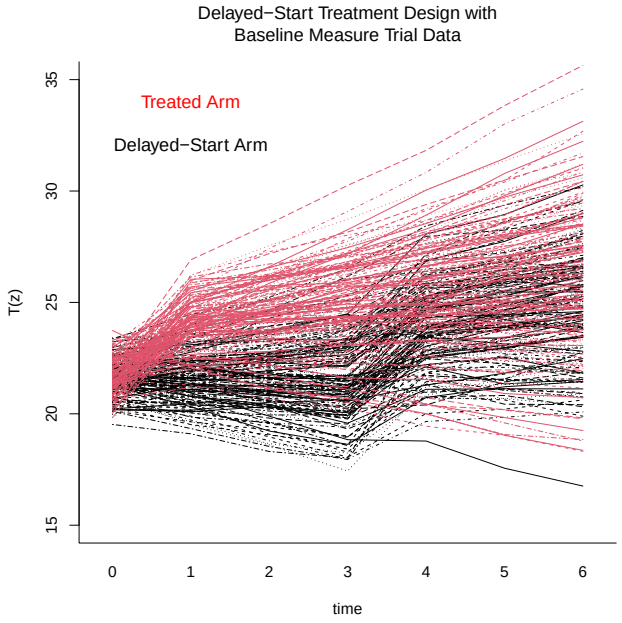
In the muscular dystrophy trial with a delayed-start of treatment, crossover from placebo into the treated group happens at time V. Notably, we do not observe data for $T(0)$ at time $t > V$ if all individuals crossover to treatment, and we assume S is measured after crossover for all individuals. In this design, we can observe both previously counterfactual values for at least some individuals, thus allowing us to estimate all correlation terms that were previously nonidentified from the data.

To be precise about when individuals receive treatment, we use the notation $T(0,t)$ and $T(1,t)$ to indicate these depend on time. Let B_i be age at randomization, τ_i be age at crossover, age_{ij} age at time j, and $\boldsymbol{t}_i$ be time from randomization so we can represent age_{ij} as $B_i + t_{ij}$. Individuals who start on treatment follow the proposed model for the $z = 1$ group. For those who crossover to treatment at age τ_i, for $age_{ij} < \tau_i$, an individual's trajectory follows the described $z = 0$ model. At the time of crossover, the trajectory of individuals changes to follow the $z = 1$ model. Our model is designed to generate data that is similar to the published natural history model by Muntoni et al. (2019). The formulation is based on reasonable expected changes based on the natural history model and has both fixed and random effects for the intercept and slope and fixed effect of age. At the time of the intervention, the subject changes their level by $(\beta_0^{(1)} + b_{i0}^{(1)} - \beta_0^{(0)} - b_{i0}^{(0)})$ and changes their slope from $(\beta_1^{(0)} + b_{i1}^{(0)})$ to $(\beta_1^{(1)} + b_{i1}^{(1)})$ from that time onward. More details and motivation can be found in Appendix J describing how Equation 3.6 is a revision of the model in Equation 3.5 using time since randomization as the time scale. Based on these modeling assumptions, we assume the mean of $\begin{pmatrix} S(1)_i & \boldsymbol{T(0,t)}_i & \boldsymbol{T(1,t)}_i \end{pmatrix}$ is

$$\begin{pmatrix} \alpha_1 + \delta_1 b_{i0}^{(0)} + \delta_2 b_{i0}^{(1)} + \delta_3 b_{i1}^{(0)} + \delta_4 b_{i1}^{(1)} \\ \beta_0^{(0)} + \beta_1^{(0)}(B_i + \boldsymbol{t}_i) + \beta_2(B_i + \boldsymbol{t}_i)^2 + b_{i0}^{(0)} + b_{i1}^{(0)}(B_i + \boldsymbol{t}_i) \\ \beta_0^{(1)} + \beta_1^{(0)} B_i + \beta_1^{(1)} \boldsymbol{t}_i + \beta_2(B_i + \boldsymbol{t}_i)^2 + b_{i0}^{(1)} + b_{i1}^{(0)} B_i + b_{i1}^{(1)} \boldsymbol{t}_i \end{pmatrix} \tag{3.6}$$

where each δ regression coefficient is a function of covariance parameters from Equation 3.4. We show an example of how these data look for crossover time $V = 3$ in Figure 3.1.

Figure 3.1: Counterfactual data for the delayed-start treatment design.



3.2.4 Conditional Independence

Conditional independence assumptions about potential outcomes are commonly used in causal inference settings to reduce or eliminate non-identified parameters and are frequently made in surrogacy validation as discussed in Conlon et al. (2017b). One such conditional independence assumption in the case of non-longitudinal T is $S(1)_i \perp T(0)_i | T(1)_i$. It was previously argued that this particular assumption was reasonable in the context where $T(1)$ is continuous and therefore

informative, particularly compared to others that could be made (Roberts et al., 2021). In our longitudinal setting, we extend these ideas of conditional independence to the random effects. To increase identifiability, we might assume $S(1)_i \perp b_i^{(0)} | b_i^{(1)}$. The consequence of this assumption in the model is that the coefficient for $b_i^{(0)}$ is equal to 0 in the $S(1)_i$ model that follows Equation 3.3. Algebraically, this is true when $\rho_T \rho_{11} = \rho_{10}$ holds. In the covariance matrix of the random slopes model in Equation 3.4, six of the ten correlations are nonidentified. For this model, we might believe some other conditional independence assumption among the measure of S and the random effects such as:

Assumption 1: $S(1)_i \perp (b_{i0}^{(0)}, b_{i1}^{(0)}) | b_{i0}^{(1)}, b_{i1}^{(1)}$. The conditional covariances $Cov(S1, b_{i0}^{(0)} | b_{i0}^{(1)}, b_{i1}^{(1)})$ and $Cov(S1, b_{i1}^{(0)} | b_{i0}^{(1)}, b_{i1}^{(1)})$ must be 0. This gives the constraints on the correlation scale that

$\rho_{1112} = \frac{\rho_{0212}^2 \rho_{S111} - \rho_{0212} \rho_{S112} \rho_{1102} + (\rho_{S102} \rho_{1102} - \rho_{S111})}{\rho_{0212} \rho_{S102} - \rho_{S112}}$ and

$\rho_{0112} = \frac{\rho_{0212}^2 \rho_{S101} - \rho_{0212} \rho_{S112} \rho_{0102} + (\rho_{S102} \rho_{0102} - \rho_{S101})}{\rho_{0212} \rho_{S102} - \rho_{S112}}$.

Assumption 2: $S(1)_i \perp b_{i1}^{(0)} | b_{i0}^{(1)}, b_{i1}^{(1)}, b_{i0}^{(0)}$ and $S(1)_i \perp b_{i0}^{(0)} | b_{i0}^{(1)}, b_{i1}^{(1)}, b_{i1}^{(0)}$. Under the multivariate normal setting, this would instead impose corresponding constraints in the form of structural zeroes on the precision matrix. While this assumption may also be plausible, we do not consider this assumption in the simulation study in this paper.

3.3 Surrogacy Validation

In the setting without longitudinal data, the CEP curve for jointly normal potential surrogate markers and outcomes has previously been shown to be $E(T(1) - T(0) | S(1) = s) = \gamma_0 + \gamma_1 s$ where γ_0 and γ_1 are functions of model parameters (Conlon et al., 2014a). This linear form and interpretation of the γ parameters as an intercept and slope term is a consequence of our normality assumptions. The CEP curve captures the validity of the surrogate through the two γ quantities.

Visually, for MVN endpoints the plot of $E(T(1) - T(0)|S(1) = s)$ over values of $S(1) = s$ for a valid surrogate would cross through the origin and have a positive slope. By construction of the principal stratification framework, a poor surrogate would fail to meet at least one criterion. We note the interpretation of γ_0 may require extrapolation when we estimate a large treatment effect on the surrogate and no data are observed in the region where $S(1) = 0$.

We explore this conditional quantity when $\boldsymbol{T}(z)$ is a vector. We derive the CEP curve when random effects exist; this requires the distribution of $\boldsymbol{T(1)} - \boldsymbol{T(0)}|S(1)$ marginalized over the random effects. Consider the simple case of random intercepts with no fixed effects of time following Equation 3.3; in this case, β_0 and β_1 are scalars. To calculate $E(T(1) - T(0)|S(1) = s)$, we integrate over the random effects: $\int_{b^{(0)}} \int_{b^{(1)}} E(T(1) - T(0)|S(1) = s, b^{(0)}, b^{(1)}) f(b^{(0)}, b^{(1)}|s) db^{(1)} db^{(0)}$. For this model, $\gamma_1 = \frac{\rho_{11}\sigma_{b1} - \rho_{10}\sigma_{b0}}{\sigma_{S1}}$ and $\gamma_0 = \beta_1 - \beta_0 - \gamma_1\alpha_1$. These quantities are similar to previous work without repeated measures.

Since $\boldsymbol{T}(0)$ and $\boldsymbol{T}(1)$ may depend on time, our validation metrics γ_0 and γ_1 can depend on time where there is, at minimum, a fixed effect of time. We denote our time-dependent curve as $CEP(t) = \gamma_0(t) + \gamma_1(t)s$. Due to the time-varying nature of the mean structure, the consequence of a fixed effect of time will be apparent in the intercept term, $\gamma_0(t)$. The slope $\gamma_1(t)$ may also depend on time when we incorporate random effects of time. While the endpoint T may be measured several times, we consider the validation measure at one time j. When we take the expected difference between $T(1)_{ij}$ and $T(0)_{ij}$ at time j from Equation 3.5, we have $\gamma_1(t_{ij}) = \frac{\sigma_{b02}\rho_{S102} + (\sigma_{b12}\rho_{S112} - \sigma_{b11}\rho_{S111})t_{ij} - \sigma_{b01}\rho_{S101}}{\sigma_{S1}}$ and $\gamma_0(t_{ij}) = \beta_0^{(1)} + \beta_1^{(1)}B_i + \beta_2^{(1)}t_{ij} - (\beta_0^{(0)} + \beta_1^{(0)}B_i + \beta_2^{(0)}t_{ij}) - \gamma_1(t_{ij})\alpha_1$. This tells us the quantity for $\gamma_1(t_{ij})$ will vary over time if there is a non-zero and non-equal covariance between $S(1)$ and the random slopes $b_1^{(0)}, b_1^{(1)}$. Further, the quantity for $\gamma_0(t_{ij})$ will depend on time if there is a non-zero and non-equal main effect of time for $\boldsymbol{T}(0)$ and $\boldsymbol{T}(1)$ outcomes (a time-treatment group interaction). Our proposed models with delayed-

start treatment give the expected difference between $T(1)_{ij}$ and $T(0)_{ij}$ and same $\gamma_1(t_{ij})$ that was derived in the standard design setting with $\gamma_0(t_{ij}) = \beta_0^{(1)} + \beta_1^{(1)} t_{ij} - (\beta_0^{(0)} + \beta_1^{(0)} t_{ij}) - \gamma_1(t_{ij})\alpha_1$.

The trial endpoint could be defined as a weighted average of the repeated measures. Rather than calculate γ_0, γ_1 at a specified time j, we could think of time continuously and calculate an average over a time range. To obtain a marginal value for overall surrogacy evaluation, we could take a weighted average of these values across a time range by integrating over a set of weights $w(t)$, $\int_t w(t) CEP(t) dt$, with different $w(t)$ such that $\int_t w(t) = 1$ based on our desired interpretation.

A consideration for conditioning and then averaging over baseline variables to calculate the CEP curve was proposed in Roberts et al. (2021) to increase plausibility of conditional independence assumptions and improve efficiency. In a clinical trial, it is usual to present marginal treatment effects, which could be estimated by calculating conditional treatment effects then averaging over the baseline covariate. If we wanted to first condition on then integrate over a baseline covariate B_i, for reasons such as making conditional independence assumptions about the outcomes conditional on these covariates, we would integrate with respect to $f(B_i|S(1) = s)$ to obtain marginal surrogacy validation quantities as detailed in the supporting information in Appendix H.

3.4 Estimation Methods for Standard Trial Design

In this section, we describe multiple approaches to estimation and inference for the model parameters and in turn for $\gamma_0(t)$ and $\gamma_1(t)$. The different methods will be compared in a simulation study.

3.4.1 Counterfactual Imputation Methods

We will assume that individuals have m measurements collected at the same time points. We also assume a standard randomized trial design with no crossover or pre-randomization values of the outcome variable. For sample size n, outcomes for each respective treatment are observed for only $n_0 = n_1 = n/2$ subjects. The complete data likelihood for this approach is the product over all n subjects of the joint distribution of the observed and counterfactual observations and the random effects. The counterfactual imputation algorithm is a Bayesian method with Markov Chain Monte Carlo (MCMC) that alternates between imputing the missing counterfactuals and corresponding random effects and drawing model parameters. The Gaussian outcomes permit conjugate priors and Gibbs sampling, though we also consider other computational options.

Let $\mu, \sigma,$ and ρ denote the mean, variance, and correlation parameters, respectively, for the model of interest; for random intercept models $\mu = \{\alpha_1, \beta_0, \beta_1\}, \sigma = \{\sigma_{S1}, \sigma_{b_0}, \sigma_{b_1}, \sigma_e\}, \rho = \{\rho_{10}, \rho_{11}, \rho_T\}$. This general notation allows for any dimension of the random effects. Conditional on parameter starting values, we impute the counterfactual values $S(1), \boldsymbol{T(0)}, \boldsymbol{T(1)}$ from the distributions of $\left(\boldsymbol{T(0)} \,\middle|\, b^{(0)}, \boldsymbol{T(1)}, b^{(1)} \atop S(1), \mu, \rho, \sigma \right)$ and $\left(S(1) \atop \boldsymbol{T(1)} \,\middle|\, b^{(0)}, \boldsymbol{T(0)}, \atop b^{(1)}, \mu, \rho, \sigma \right)$.

From these, we draw the random effect estimates $b_{\cdot}^{(0)}, b_{\cdot}^{(1)} | S(1), T(0), T(1), \mu, \rho, \sigma$ from a multivariate normal distribution. More details on deriving conditional forms of the distribution to impute counterfactuals and drawing random effects can be found in the supporting information in Appendices G and H. We assume vague priors for the identified mean parameters μ and inverse gamma priors for the variances σ^2. Conditional on the imputed outcomes, $b_{\cdot}^{(z)}$, μ, and σ, we draw the correlation parameters ρ. To control the prior distribution on each correlation term, we decompose the variance-covariance matrices Ψ into standard deviations S and correlations R (Barnard, McCulloch, and Meng, 2000). In many cases, we assume Uniform(-1, +1) priors and compare

this to more informative rescaled Beta priors placed on nonidentified correlations when we have reason to expect the correlation is likely to be non-negative. For computational efficiency, we use a Metropolis-Hastings (MH) step with Fisher's z-approximation and Jacobian transformation as described in Appendix I. The positive definiteness constraint on R is based on the determinant being positive: for the random intercept model, this is analytically written as $1 - \rho_T^2 + \rho_{11}^2 > \rho_{10}^2$. For the random slope, the closed form boundaries for each correlation parameter are more complex to solve, and the positive definiteness constraint is handled by rejection. Interestingly, the positive definite constraint for the random intercept model is fulfilled when the conditional independence constraint is assumed.

To further help with identifiability and efficiency, we can implement our proposed conditional independence assumptions. For the random intercept conditional independence assumption and assumption 1 for random slopes, we fulfill the constraints using algebraic equalities. To implement the second conditional assumption for random slopes, we could use the precision matrix decomposition described in Wong, Carter, and Kohn (2003) since there are two structural zeroes in Ψ^{-1} when this assumption holds. Once all parameters are drawn, we calculate γ_0, γ_1 based on the formulas provided in Section 3.3 and repeat this process over many iterations to obtain posterior distributions.

3.4.2 Observed Data Methods with Random Effects

Alternatively, we can use only the observed data likelihood without imputing counterfactuals, where the observed data likelihood is comprised of the observed values of S and T and the corresponding random effects. For example, the observed data for m time points are $T(0)_{l1}, T(0)_{l2}, ..., T(0)_{lm}$ for $l = 1, ..., n_0$ and $S(1)_k, T(1)_{k1}, T(1)_{k2}, ..., T(1)_{km}$ for $k = 1, ..., n_1$. Consider the random intercept model where $T(0)_{lj} = \beta_0^{(0)} + b_l^{(0)} + e_{lj}$ and $T(1)_{kj} = \beta_0^{(1)} + b_k^{(1)} + e_{kj}$

with $e_{\cdot} \sim N(0, \sigma_e^2)$. From the distribution of $(S(1)_i \;\; b_i^{(0)} \;\; b_i^{(1)})^T$ in Section 3.4.1, we need only the components $b_l^{(0)} \sim N(0, \sigma_{b_0}^2)$ and $\begin{pmatrix} S(1)_k \\ b_k^{(1)} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{S1}^2 & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ & \sigma_{b_1}^2 \end{pmatrix}\right)$ that correspond to observed data.

For this simplified algorithm, we estimate the identified parameters and $b^{(0)}, b^{(1)}$ related to $S(1), T(1)$ for those with $z = 1$ separately from those for $T(0)$ with $z = 0$, and we use the same priors described in the previous section, namely normal priors on the means $(\alpha_1, \beta_0, \beta_1)$, inverse-gamma priors on the variance terms $(\sigma_{S1}, \sigma_{b_1}, \sigma_{b_0}, \sigma_e)$, and either Uniform or rescaled Beta priors on the correlations $(\rho_{11}, \rho_T, \rho_{10})$ for random intercept models. Since σ_e is shared between the two arms of the study, it is drawn based on the residuals of both arms. While ρ_{11} is identified and drawn based on Fisher's z-approximation from data for the $z = 1$ arm, the nonidentified correlation ρ_T is drawn from its prior, and conditional independence can be assumed to solve for ρ_{10}. We directly calculate γ_1, γ_0 from these quantities without the imputation of the counterfactual outcomes.We implement this method using MCMC, though other non-Bayesian options are available, and this algorithm is applicable to more complex mean structures and random slopes models.

3.4.3 Observed Data Methods Integrating Out Random Effects

For this Gaussian model, we can also integrate out $b_l^{(0)}, b_k^{(1)}$ from the likelihood and calculate γ_0 and γ_1 directly. In addition to using the observed data without imputing counterfactuals as described in Section 3.4.2, we can maximize the likelihood or posterior directly and bypass the estimation of random effects. The random intercept model for this method can be written out for m time points from the marginal (meaning integrated over the random effects) likelihood, where for $z = 0$, the

vector $\boldsymbol{T(0)}$ has mean $\boldsymbol{\beta}_0^{(0)}$ and covariance $\begin{pmatrix} \sigma_e^2+\sigma_{b_0}^2 & \cdots & \sigma_{b_0}^2 \\ & \ddots & \vdots \\ & & \sigma_e^2+\sigma_{b_0}^2 \end{pmatrix}_{m\times m}$.

Similarly for $z = 1$, $(S(1) \quad \boldsymbol{T(1)})$ has mean $(\alpha_1 \quad \boldsymbol{\beta}_0^{(1)})$ and covariance $\begin{pmatrix} \sigma_{S1}^2 & \rho_{11}\sigma_{S1}\sigma_{b1} & \cdots & \rho_{11}\sigma_{S1}\sigma_{b1} \\ & \sigma_e^2+\sigma_{b_1}^2 & \cdots & \sigma_{b_1}^2 \\ & & \ddots & \vdots \\ & & & \sigma_e^2+\sigma_{b_1}^2 \end{pmatrix}_{m+1\times m+1}$.

For comparability across methods, the prior distributions proposed earlier can be used to draw ρ_{10} and ρ_T, and in turn calculate γ_0 and γ_1. Standard errors can be derived in part from the Hessian matrix. We draw ρ_T from a Beta distribution 100 times, use the mean of the draws for the point estimate of ρ_T, and account for this variability in the final γ variability estimates using a within- plus between-variance type formula (detailed in Appendix L). Since this method does not require the computation of repeated draws of random effects, this option lends itself well to a maximum likelihood-based method and sensitivity analyses of fixing the nonidentified parameters to a range of values. For this optimization method, we suggest maximizing the likelihood using the `optim` function in `R` (R Core Team, 2012). The procedure is also applicable to random slope models, though the closed-form, marginal distribution is more complex and not included here.

3.4.4 Methods Simplifying the Repeated Trial Outcome

Whereas the previous sections considered models and algorithms for the repeated measures, we consider alternative ways to simplify these data. Albert (1999) describes standard options to summarize longitudinal observations for a clinical trial, which will be similar to options for summarizing a longitudinal metric of surrogacy. For efficiency comparisons, we simply use the first

measurement or average of the repeated measurements and define $T(0)_1, T(1)_1$ or $\overline{T(0)}, \overline{T(1)}$ as the outcomes respectively. By condensing the outcomes into scalars, our algorithm is the same as when we did not consider repeated measurements (Roberts et al., 2021). We compare these two methods to fitting mixed models on the repeated measurements.

3.5 Simulation Studies

The purpose of our simulation study is to mimic clinical trials. In the standard randomized design, we have non-identified parameters. In designs with pretreatment measures or delayed-treatment start for the placebo arm, some non-identifiability is alleviated. We compare the performance of estimation strategies proposed in Section 3.4 of estimating the γ quantities of interest. We also assess the robustness of our results across sample sizes and model misspecification.

3.5.1 Results Comparing Methods and Models with Random Intercepts

We compare algorithmic and modeling strategies for the random intercept model for 200 simulation replications. In this section, there are three repeated $T(0), T(1)$ outcomes and one cross-sectional, normally-distributed value of S that is correlated with the random effects for $T(z)$. Data are generated with $n = 300$ subjects, with true values: $\alpha_1 = 2, \beta_0 = 22, \beta_1 = 23.1, \rho_{10} = 0.15, \rho_{11} = 0.7, \rho_T = 0.214, \sigma_{S1} = \sigma_{b0} = \sigma_{b1} = 0.25, \sigma_e = 0.3$ (Equation 3.2 and 3.3).

In fitting the models, we use either a Beta$(8, 5)$ prior that is rescaled, meaning we draw $U \sim Beta(8, 5)$ then transform the draw to $x = 2 \times U - 1$, or a Uniform prior on the parameter ρ_T. We note that this rescaled Beta(8,5) prior has mean equal to 0.231 and standard deviation equal to 0.260 and is therefore mildly informative. The observed data scenarios explained in Section 3.4.2 are denoted *Obs Data*, whereas rows marked *Imputation* use the imputation algorithm in

4.1. Both algorithms are consistent with the generated data, since the differences between the two algorithms are primarily in how the non-identified parameters and counterfactual outcomes are handled. Lastly, we vary making the conditional independence assumption across methods, which is denoted as *CI* or *No CI*. The likelihood-based optimization method from Section 3.4.3 is labeled as such. The simplified methods considered in Section 3.4.4 are shown in the bottom two rows of the results table.

Notably, the concept of true, data generating values and corresponding bias of the estimates may be considered somewhat ambiguous when handling nonidentified parameters. To account for the fact that several sets of values for the non-identified parameters could generate the observed data, we take a more broad approach of defining the 'true' values of γ_0 and γ_1. Similar ideas are explored by Zhang and Rubin (2003) by creating large sample bounds of causal quantities; these tend to be quite wide in practice. In our work, the range of values that could generate the simulated data are determined by obtaining thousands of draws from a non-informative prior distribution for non-identified parameters, fixed values of the identified parameters, and evaluating which sets of correlations produce positive definite covariance matrices. These can be thought of as possible generative matrices in the infinite data case where we consider a non-informative prior to be Uniform(-1, 1), and we describe more motivation and complexities of this procedure in Appendix K. The corresponding, valid γ values describe a range of possible truths.

The simulation results for γ_0 and γ_1, found in Table 3.1, show that in the setting of non-identified parameters, the standard error estimates are larger than the standard deviation of the point estimates across simulations. The impact of different priors on ρ_T is demonstrated between the first set of results. As we would expect in the setting of non-identified parameters, there is some sensitivity to the prior, particularly in the standard error estimates. When holding all else constant, we see that assuming conditional independence results in gains in efficiency. When comparing the impact of

different algorithms for estimating random effects, we expect the row of results involving imputation of counterfactuals to match the results for using the observed data without imputing counterfactuals. This suggests we can implement the more efficient algorithm using the observed data only and expect similar results to using the complete likelihood-based method. The method that maximizes the observed data likelihood provides similar results to the other observed data methods, demonstrating its potential as a computationally efficient alternative to the Bayesian methods. We see some gains in efficiency when we use the average of three measures compared to using only the first measurement when comparing the standard deviation of point estimates for γ_0 and γ_1. Notably the mixed modeling approach, which makes full use of the available information in the data, was most efficient. Across different modeling assumptions, some of the simulation results largely suggest that S would be a valid surrogate based on the estimated values of γ_0 and γ_1. In particular, as desired for a valid surrogate, the credible interval for γ_0 covers 0 while the credible interval for γ_1 does not in the settings where informative priors (Beta) and conditional independence are assumed. This underscores the importance of implementing context-plausible modeling assumptions.

We assess robustness by varying the sample size in simulations. In results shown in Appendix M, we see that for sample sizes considered (100 - 1,000), the average standard error is larger than the standard deviation of the point estimates, as non-identifiability persists even with increasing sample size. Noting data from only $n/2$ subjects will be available for estimation using the observed data algorithm, smaller sample sizes result in noticeably larger of variance.

Model misspecification is likely in practice. To explore the consequence of model violation, we generated the joint distribution of the random effects and $S(1)$ as skewed based on a correlated multivariate Gamma distribution with means adjusted to be $(\alpha_1, 0, 0)$. Based on results found in Appendix M, we see differences in estimation properties. The average standard error across

simulated datasets is larger for the skewed, misspecified distribution of the random effects. The average posterior means are also different, though they fall within the true, generated values. From this, we conclude that there will be differences in estimated γ_0 and γ_1 quantities when models are misspecified, though the algorithm is still able to be fit and produces reasonable results.

3.5.2 Results Comparing Assumptions with Random Slopes

We conduct simulations with 200 replications for more complex random slope models in a larger trial size of $n = 900$ to assess the performance. There are six repeated measurements of T post-randomization for both arms. We generate data according to Equation 3.5 with generating values $\alpha_1 = 2, \beta_0^{(0)} = 22, \beta_0^{(1)} = 23.45, \beta_1^{(0)} = 0, \beta_1^{(1)} = -0.4, \beta_2 = 0, \rho_{S01} = 0.025, \rho_{S11} = 0.022, \rho_{S02} = 0.71, \rho_{S12} = 0.25, \rho_{0111} = 0.15, \rho_{0102} = 0.03, \rho_{0112} = 0.05, \rho_{1102} = 0.02, \rho_{1112} = 0.10, \rho_{0212} = 0.25, \sigma_{S1} = \sigma_{b01} = \sigma_{b02} = 0.75, \sigma_{b11} = \sigma_{b12} = 0.45, \sigma_e = 0.15$. We use the Bayesian observed data algorithm described in Section 3.4.2 as we expect these results to mimic those from an imputation method. The metrics are shown at various time points to demonstrate their time-varying nature. Again we provide the true values of γ based on infinite data scenarios as described in Section 3.5.1 since the observed data in the simulations may be consistent with multiple full data likelihoods.

Our results are shown in Table 3.2, and we compare estimation properties when we do or do not make the first conditional independence assumption. The results suggest that the surrogate might be valid at the first two evaluated time points (1-2) based on the credible interval for γ_0 containing zero and the interval for γ_1 being strictly positive. We again see gains in efficiency when we make the conditional independence assumption. The credible intervals also become wider as time increases, which is expected based on the increasing complexity of formulas for γ_0 and γ_1 over time.

3.5.3 Results for Pretreatment Measures

We consider the setting of a pre-treatment measurement of T and six repeated measurements post-randomization. We contrast modeling this measure as an outcome in the mixed model versus as a baseline covariate. The simulations use the observed data only, so we can further compare this to the previous section where we do not incorporate this measurement. To obtain the presented marginal quantities over the baseline covariate for γ_0 and γ_1, we complete a marginalization step as described in Section 3.3 and Appendix H. The main results, shown in Table 3.3 part a, display the estimated surrogacy measures at time one. The results incorporating this pretreatment measure show an efficiency gain by including $T(0)$ at baseline in the model as an outcome. We would recommend using this measurement as an outcome in the model, as it is also more computationally efficient than using it as a covariate and integrating over it. One reason to use $T(0)$ at baseline as a covariate is to make the conditional independence assumptions more plausible.

3.6 Motivating Data Example

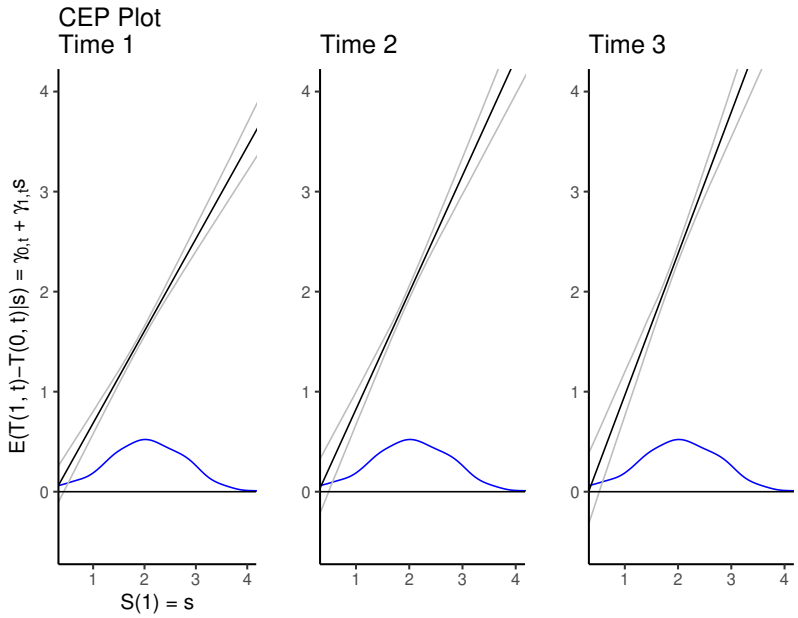
The motivating study for the proposed method investigates a gene therapy for Duchenne muscular dystrophy patients. This method was developed to accommodate the particularities of the study design and longitudinal data collection. Based on published literature related to the disease, we propose to model effects over time since age, growth, and disease deterioration have strong effects on the functional outcome of interest, North Star Ambulatory Assessment (NSAA) score, T. This trial is ongoing and has not been unblinded, so individual patient level data is not publicly available. The presented data are simulated from parameter values to match summary statistics from the preliminary data or natural history plots when available. As estimates of random effects were not available, the authors chose parameters to match what they considered reasonable values. Briefly,

we consider the delayed-treatment design with longitudinal measures of the outcome T and a post-treatment measure of gene expression S. Data include a crossover from placebo to treatment group at time $V = 3$ with a total of $k = 6$ post-treatment measurements and a measure of T at baseline. Based on the model in Eq. 6 and a trial with sample size $n = 900$, we use the generating values described in section 5.2 with details on the procedure found in Appendix J.

Based on these data, we estimate the parameters using the observed data algorithm. First we look at tabular results across multiple datasets for γ evaluated at time one in Table 3.3 part *b*. Overall, there are large efficiency gains in the standard error when using the data available from a delayed-treatment start design compared to the other considered designs that do not. There are further efficiency gains when there is a pre-treatment measurement of T as well. Notably, the average standard error is close to the standard deviation of the point estimates for the delayed-treatment start setting, because we have identifiability. This substantial reduction in the standard error demonstrates the clear gains of conducting a delay-start design from a statistical perspective.

For one data set, we plot the corresponding CEP curves in Figure 3.2 since the $\gamma(t_{ij})$ quantities are a function of time. From these CEP curves, we see that there is some treatment-time interaction based on changing $\gamma(t_{ij})$ estimates across evaluation times. S may be a valid surrogate at any time point where the credible interval for γ_0 overlaps 0 and the interval for γ_1 does not, which occurs at times one through three for this dataset. Importantly, we see the density of $S(1)$ in Figure 3.2 and are reassured that the values of $E(T(1,t) - T(0,t)|S(1) = s)$ are positive where we observe data for $S(1)$, indicating proper surrogacy should hold. In other trials, it is plausible that S may only be valid closer to the time that T is measured or that its value is less clear in the region where $S(1)$ is observed. We see in the differences between Table 3.3 part a and b that assessment of surrogacy would be more difficult without the data collected from the delayed-start treatment design.

Figure 3.2: Results of the CEP plot where age is both a fixed and a random effect for the delayed-treatment Muscular Dystrophy trial simulated data with baseline measures of T. Each panel denotes a different time point that the surrogate is evaluated. The displayed credible intervals are pointwise credible intervals of $\gamma_0 + \gamma_1 s$. The blue curve shows the density plot of observed values of S.



We explored the consequences of mean structure specifications that differ from the generating model. Here we investigated when the data exhibits a quadratic effect of time, but the model only accounts for a linear effect of time and see larger variability of the estimates in this case. We anticipate that model selection could be performed in this setting to accurately specify fixed effects.

3.7 Discussion and Future Directions

We have proposed a causal inference approach to validating surrogate endpoints when the true outcome of interest is measured repeatedly throughout the trial. This paper demonstrates several algorithms and benefits of incorporating the mixed modeling framework, particularly in the set-

ting of a delayed-start treatment design. In our longitudinal setting, the potential for crossover treatment arms or use of a pretreatment measurement allows us to estimate potentially identified correlation parameters that arise between treatment arms, indicating benefits of this trial design. Beyond the proposed model, we could also let the quantities γ_0 and γ_1 depend on B_i by allowing the coefficients for B_i to differ between treatment arms, or we could include age-treatment, age-squared-treatment, or age-surrogate interactions in the model, for example. There is potential for efficiency gains when incorporating the proposed conditional independence assumptions of the random effects.

In Section 3.4, we introduced three alternatives for estimation in the general longitudinal setting. While all of the algorithms should produce similar results, we suggest using the observed data methods for computational gains. In particular, the likelihood based method will be the fastest. However, parameterizing the longitudinal model after marginalizing over the random effects can be challenging when random slopes are modeled, and the variance parameters may be poorly estimated in some settings. The proposed optimization method that marginalizes over random effects will also be more difficult in complex situations such as generalized linear mixed models where an approximation of integrals may be necessary. For these reasons, we believe that the Bayesian methods using MCMC with the observed data likelihood and prior distributions are a reliable choice.

Further questions can be explored in this data setting regarding trial design. It is important to understand the amount of efficiency gain that is possible by including the baseline NSAA score. Similarly, it is beneficial to quantify the gains of conducting the crossover portion of the trial and assess the optimal time to do crossover. This must be considered within the cost-benefit context of treating all patients in the placebo arm and following them over time. Given our observation that small sample sizes can result in wide credible intervals of the validation quantities, it is of interest

to explore how historical trials may be leveraged when the current trial has a small sample size.

As noted by Vanderweele (2013), additional criteria (one-sided average causal sufficiency) may assure against the so-called surrogate paradox, an unfortunate phenomenon that can occur when the treatment effect on the surrogate is positive, the correlation between the surrogate and true outcome is positive, yet the treatment effect on the true outcome is negative. Further work in this area is available in Elliott et al. (2015) and Price et al. (2018). A limitation of the proposed method is that it relies on a properly specified, parametric model to validate surrogate endpoints. The model can be extended to be more flexible, handle more types of endpoints, or accommodate time-varying covariates (see Kim et al., 2017 for an example). As noted in previous work, the CEP curve framework does not directly rely on the normality assumption, and copula or other modeling could be implemented instead (see Taylor, Conlon, and Elliott, 2015). This becomes more complex with the repeated measures, but extensions into generalized linear mixed models would be an interesting area of future research. Further, the time-varying CEP curve will be relevant for time-to-event data where censoring and semi-competing risks must be taken into account.

3.8 Data Availability and Software

Due to the proprietary and ongoing nature of the clinical trial, the individual patient level data is not available for use. Programming was done in R v3.6.2 (R Core Team). The R code for illustrative simulation studies is available at https://github.com/emilykroberts/Surrogacy-Validation-Longitudinal-Outcomes. The content of this chapter has been accepted for publication in *Biometrics*.

3.9 Tables

Setting	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
Range of Data Generating Value Unif, No CI	(-2.052, 1.466)			(-0.182, 1.575)		
Comparing the Impact of Priors and CI Assumptions						
MCMC, Obs Data, Random Effects, No CI, MH Unif(-1, 1) Prior	-0.300	0.931	0.143	0.706	0.470	0.072
MCMC, Obs Data, Random Effects, CI, MH Unif(-1, 1) Prior	-0.264	0.736	0.174	0.686	0.372	0.087
MCMC, Obs Data, Random Effects, No CI, MH Beta(8, 5) Prior	-0.057	0.914	0.160	0.581	0.461	0.080
MCMC, Obs Data, Random Effects, CI, MH Beta(8, 5) Prior	0.002	0.408	0.155	0.552	0.205	0.078
Comparing the Impact of Different Methods/Algorithms						
MCMC, Imputation, Random Effects, CI, MH Beta(8, 5) Prior	0.022	0.392	0.161	0.542	0.199	0.080
MCMC, Obs Data, Random Effects, CI, MH Beta(8, 5) Prior	0.002	0.408	0.155	0.552	0.205	0.078
Optimization, Obs Data, Integrated, CI, MH Beta(8, 5) Prior	0.046	0.404	0.150	0.528	0.201	0.074
MCMC, Obs Data, First Measure, CI, Beta(8, 5) Prior	0.038	0.441	0.198	0.530	0.211	0.095
MCMC, Obs Data, Measure Average, CI, Beta(8, 5) Prior	0.052	0.448	0.152	0.522	0.212	0.072

Table 3.1: Simulation results of random intercept models comparing different assumptions and models. The true values of γ_0 and γ_1 are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices under conservative settings, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a Uniform(-1, 1) distribution with no conditional independence (CI) assumptions. Results shown for γ quantities are the posterior mean, average estimated standard error within simulation, and standard deviation of the point estimates across simulation replications. Obs Data represents algorithms that use only the observed data (rather than an imputation scheme), and MH denotes when Metropolis Hastings steps were involved in the Markov Chain Monte Carlo (MCMC).

	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
Range for Data Generating Values at Time 1	(-1.508, 1.356)			(0.145, 1.577)		
No CI	0.029	0.789	0.093	0.807	0.395	0.043
CI	-0.050	0.612	0.103	0.848	0.306	0.045
Range for Data Generating Values at Time 2	(-1.900, 1.945)			(0.048, 1.973)		
No CI	0.224	1.093	0.122	0.908	0.547	0.056
CI	0.090	0.832	0.147	0.978	0.416	0.068
Range for Data Generating Values at Time 3	(-2.399, 2.647)			(-0.101, 2.422)		
No CI	0.433	1.427	0.159	1.003	0.714	0.071
CI	0.231	1.084	0.201	1.108	0.541	0.094
Range for Data Generating Values at Time 4	(-2.943, 3.392)			(-0.273, 2.894)		
No CI	0.642	1.774	0.199	1.099	0.888	0.089
CI	0.372	1.349	0.258	1.238	0.674	0.121

Table 3.2: Simulation results of random slope models over time. The true values of γ_0 and γ_1 are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a non-informative prior. CI or No CI denotes whether conditional independence was assumed.

	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
a						
Range for Data Generating Values at Time 1	(-1.508, 1.356)			(0.145, 1.577)		
Standard Design, No T_{BL} Covariate, No CI	0.029	0.789	0.093	0.807	0.395	0.043
Standard Design, No T_{BL} Covariate, CI	-0.050	0.612	0.103	0.848	0.306	0.045
Standard Design, T_{BL} As Outcome, No CI	0.053	0.466	0.153	0.776	0.233	0.059
Standard Design, T_{BL} As Outcome, CI	0.015	0.324	0.111	0.802	0.165	0.053
Standard Design, T_{BL} As Covariate, No CI	-0.037	0.764	0.123	0.835	0.388	0.055
Standard Design, T_{BL} As Covariate, CI	-0.103	0.614	0.109	0.870	0.313	0.050
b						
Delayed-Start Design, T_{BL} As Covariate, No CI	-0.012	0.149	0.141	0.868	0.071	0.066
Delayed-Start Design, T_{BL} As Covariate, CI	-0.038	0.126	0.093	0.837	0.060	0.048

Table 3.3: Simulation results of random slope models comparing trial designs. Here we compare either a standard randomized design (*a*) to a delayed-start treatment design (*b*). We also assess the impact of incorporating a pre-treatment, baseline measurement of T, T_{BL}. In some scenarios, T_{BL} is treated as a covariate and in others as an outcome in the mixed models. The true values of γ_0 and γ_1 are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a non-informative prior. CI or No CI denotes whether conditional independence was assumed.

CHAPTER IV

Surrogacy Validation for Time-to-Event Outcomes with Illness-Death Frailty Models

4.1 Introduction

Time-to-event endpoints are common in oncology trials, though it can often take many years to accrue enough observed events to complete the study (Kemp et al. 2017). In a randomized clinical trial, an appropriate surrogate endpoint can serve as a substitute indicator for if a treatment effect exists on some true outcome of interest. In this work, our data come from a prostate cancer clinical trial with a binary treatment of adding anti-androgen therapy to an existing regimen (Shipley et al. 2017). The two endpoints of interest are the occurrence of distant metastasis and overall survival. Here the terminal event is death from any cause and is the primary endpoint for the trial. For these patients, death from prostate cancer will only occur if the person has had metastases. However, some men will experience death during follow-up with or without experiencing distant metastases spreading first. Overall survival is therefore a mixture of two death types, death from prostate cancer and death from other causes. However, in the data the cause of death is not known. Mechanistically understanding whether distant metastases is a desirable surrogate for overall survival in this setting may be beneficial for clinicians and trialists.

Given the substantial risk of potentially using an invalid surrogate endpoint in a large scale trial, rigorous standards have been proposed to validate a surrogate (Vanderweele, 2013). The first criteria to determine the validity of candidate surrogate endpoints were suggested by Prentice (1989) which test whether a treatment affects the true endpoint only through the pathway of the surrogate endpoint. While the criteria are applicable to different outcomes such as time-to-event endpoints that we will be focusing on, they involve regression models that rely on conditioning on the observed value of S, leading to a non-causal interpretation. More recent frameworks to determine if a surrogate marker is appropriate for use in a future trial can be broadly grouped into the causal effects and causal association paradigms (Joffe and Greene, 2009). The causal association framework aims to evaluate the relationship of the treatment effect on the surrogate S with the treatment effect on the true clinical endpoint T. These methods are often built upon counterfactual outcomes $T(z)$, which are the clinical outcomes of interest, and $S(z)$, the surrogate endpoints, where the notation $Z = z$ represents the treatment under either the observed or counterfactual assignment.

Methods within the causal association framework have been proposed for trials where the true outcome T is a time-to-event outcome under different corresponding surrogate endpoint types. Tanaka et al. (2017) consider a binary surrogate for a survival primary outcome within the meta-analytic framework, and Gao (2012) considers a time-to-event T and binary S for a single trial using principal stratification methods (Frangakis and Rubin, 2002). Taylor et al. (2015) propose a Gaussian copula model with a survival endpoint for T and ordinal endpoint S. The principal stratification estimand proposed by Qin et al. (2008) allows for a continuous S and time-to-event T. This was expanded upon in Gabriel and Gilbert (2014) and Gabriel, Sachs, and Gilbert (2015) in pursuit of a causal effect interpretation. Causal solutions for validation become more challenging when the surrogate is also subject to censoring. Instead, others such as Parast and colleagues (2017) rely on different measures such as proportion explained for time-to-event outcomes, and

likewise Hsu et al. (2015), Vandenberghe et al. (2018), and Weir et al. (2021) address time-varying surrogates using mediation approaches that rely on proportion mediated metrics within the causal effects paradigm.

To our knowledge, the setting where both S and T are time-to-event endpoints has not been fully addressed within the principal stratification framework. Building on the work of Frangakis and Rubin (2002), we aim to develop a corresponding Causal Effect Predictiveness (CEP) curve proposed by Gilbert and Hudgens (2008) to validate a surrogate endpoint when both S and T are time-to-event. The key to obtaining a causal assessment in this paradigm is classifying individuals based on their set of potential values of the post-treatment variable, which here would be the surrogate endpoint. In a simple case where S and T are Gaussian outcomes and Z takes on the value 0 or 1, the analog to surrogate-specific strata and the corresponding CEP curve for validation is based on the quantity $E(T(1)-T(0)|S(1)-S(0) = s)$. Briefly, the CEP criteria intuitively assert that there be no average treatment effect on T for the strata of patients defined by no treatment effect on S, and conversely that there exist an overall treatment effect on T for the strata of patients defined by a treatment effect on S. A comparable contrast and consideration of principal strata when $T(z)$ and $S(z)$ are subject to censoring and a semi-competing risk structure will be explored in this chapter.

Outside of the surrogacy validation setting, semi-competing risks based on counterfactual hazards have been explored (Huang, 2021). Within the principal stratification framework, unobserved outcomes due to truncation by death can be addressed by defining strata based on survivorship cohorts (Zhang and Rubin, 2003). Comment et al. (2019) define a survivor average causal effect in the presence of a semi-competing risk where principal causal effects are defined for individuals who would survive regardless of the assigned treatment. Xu et al. (2020) propose a causal estimand for a semi-competing risk structure to address truncation by death $\frac{P(S(1)<\tau|T(0)\geq\tau,T(1)\geq\tau)}{P(S(0)<\tau|T(0)\geq\tau,T(1)\geq\tau)}$

which conditions on these survivor principal strata.

The estimands for surrogacy validation with a continuous S by Qin et al. (2008) and Gabriel, Sachs, and Gilbert (2015) described earlier can be written as $1 - \frac{P(T(1)=\tau|T(1)\geq\tau_{k-1},S(1)=s_1,S(0)=s_0)}{P(T(0)=\tau|T(0)\geq\tau_{k-1},S(1)=s_1,S(0)=s_0)}$ and $\frac{1-P(T(1)>t|T(0)\geq\tau,T(1)\geq\tau,S(1)=s_1,S(0)=s_0)}{1-P(T(0)>t|T(0)\geq\tau,T(1)\geq\tau,S(1)=s_1,S(0)=s_0)}$ for some time τ, respectively. In our setting where S may not be observed before T, our goal of conditioning on counterfactual surrogate outcomes as suggested by the previous CEP quantities becomes less straightforward while accounting for semi-competing risks. For example, while it may be possible to condition on strata defined by $S(0)$ and $S(1)$ occurring or not by time τ, the proper surrogacy validation estimand remains unclear. For example, candidate estimands may include either $\frac{P(T(1)<\tau|S(0)\geq\tau,S(1)\geq\tau)}{P(T(0)<\tau|S(0)\geq\tau,S(1)\geq\tau)}$ or $\frac{P(T(1)<\tau|T(0)\geq\tau,S(0)\geq\tau,S(1)\geq\tau)}{P(T(0)<\tau|T(1)\geq\tau,S(0)\geq\tau,S(1)\geq\tau)}$ for some time τ.

Rather than conditioning on surrogate outcomes, we develop a principal stratification approach that conditions on counterfactual hazards and outline causal quantities based on these. We propose an illness-death model to incorporate the censored and semi-competing risk structure of the data. Previous work using principal surrogacy for repeated outcome measurements incorporates estimation of subject-specific random effects in Chapter III. Here we utilize frailty terms to capture subject specific heterogeneity and allow dependence among the transitions of the illness-death model. Frailties have been proposed for surrogate validation settings that differ from our single trial with subject-level, counterfactual outcomes. These methods include joint frailty-copula models for meta-analysis to define valid surrogates (Emura et al., 2017; Sofeau, Emura, and Rondeau, 2019; Sofeau, Emura, and Rondeau, 2020).

In Section 4.2, we propose the causal modeling strategy based on the illness-death approach for a single trial and link this formulation to the Prentice criteria. In Section 4.3, we provide the likelihood of the illness-death model and propose a Bayesian estimation strategy. Section 4.4 describes our proposed CEP quantities and explores CEP plots that correspond to different data

settings to help define what an ideal surrogate would look like. A simulation study is provided in Section 4.5 with a real data analysis from a prostate cancer trial in Section 4.6. Discussion and future work are provided in Section 4.7.

4.2 Illness-Death Approach

The structure of the illness-death model is a natural way to describe data with the semi-competing risk structure and has potential use for surrogacy validation (O'Quigley and Flandre, 2012). Here we consider counterfactual illness-death models and the principal stratification framework. Let $T_{jk}(z)$ denote the gap time between two states $(j = 1, 2, k = 2, 3)$ and corresponding transition intensities λ^z_{jk} between states in the treatment-specific illness-death models for treatment $Z = z$ as shown in Figure 4.1.

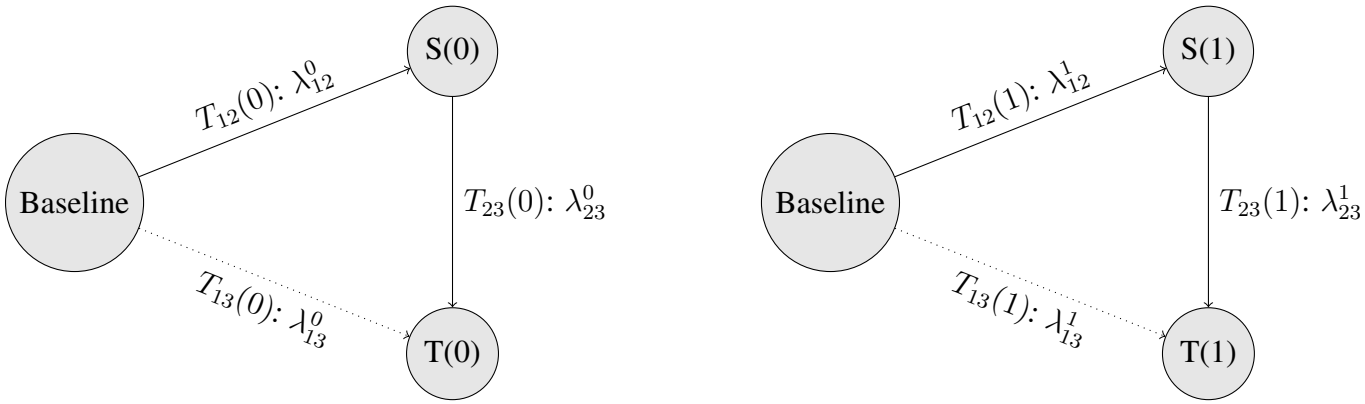


Figure 4.1: Counterfactual illness-death models for baseline, illness (S), and death (T). The potential pathways are labeled with the gap time and corresponding transition intensity for each treatment arm.

Notably, this conceptualization is related to the models used in the Prentice criteria (1989). In short, the Prentice criteria assess whether a) the treatment and true endpoint are conditionally independent, given the surrogate endpoint, and b) the surrogate and the treatment are correlated. This determination is made by fitting two regression models and determining if the coefficient for

the treatment effect becomes null after adjusting for the surrogate in the model. These ensure that a treatment effect on the true endpoint will imply a treatment effect on the surrogate endpoint. In particular, Prentice's measures, which identify statistical surrogates, are only correlative.

We propose a more rigorous and flexible strategy to identify a consistent surrogate using potential outcomes and counterfactual illness-death models in pursuit of a causal interpretation (VanderWeele, 2013). Motivation for our proposed models can be seen through a special case of regression models that are related to models used to evaluate the Prentice criteria. For example, consider the models for the observed data

$$\lambda_{12}(t)\exp(\omega_{12i} + \phi_1\, Z_i + \eta_1\, X_i) \tag{4.1}$$

$$\lambda_{13}(t)\exp(\omega_{13i} + \phi_2\, Z_i + \eta_2\, X_i)$$

$$\lambda_{23}(t)\exp(\omega_{23i} + \theta\, S_i + \phi_3\, Z_i + \eta_3\, X_i + \beta\, S_i\, Z_i)$$

where S denotes the time of the surrogate outcome occurring, ω_{jk} denote frailty terms, Z denotes treatment, X denotes baseline covariates, and t is measured from randomization. These three models can be viewed as a generalization of the models suggested by Prentice. Some differences with our proposed models are that they have additive transitions, allow for more interaction terms (when $\beta \neq 0, \phi_1 \neq \phi_2 \neq \phi_3 \neq 0$) and include frailties. Further extension of the models and their connection with the counterfactual illness-death models in Figure 4.1 can be found in Appendix O. In the model we propose and explore in detail in the following sections, each counterfactual arm has its own set of transition hazard models.

4.2.1 Defining Causal Quantities Based on Hazards and Frailty Models

We propose to model the transition intensities that correspond to the gap times $T_{jk}(z)$ in Figure 4.1. Shared or common frailty terms, which quantify the dependence between the different processes within the same person, can provide information on the dependence structure between the time to intermediate event and the time to terminating event in standard multi-state models (Zhang et al., 2014; Xu et al., 2010). Frailties are commonly incorporated to model correlation among events, heterogeneity among individuals, or to capture the effect of some omitted covariate. In our setting, we consider both counterfactual outcomes and transitions, and we want to allow for possible dependence between the counterfactual outcomes. As this association is integral to the value of the surrogate, we propose to use illness-death frailty models where the hazards are linked via frailty terms. Here we consider multiple hazards with frailties both to allow dependence across state transitions and to link observable transitions in arm $Z = z$ to the counterfactual transitions for $Z = 1 - z$.

For a single time-to-event and a general frailty ω, the hazard can be written $\lambda(t|X, \beta, \omega, \kappa) = \lambda_0(t) \exp(\kappa\omega + X\beta)$, where ω has some pre-specified distribution and may have an associated coefficient parameter κ. Various assumptions can be made about the frailty term ω, such as that it follows a Normal or Gamma distribution, for simplicity and computational feasibility. For the illness-death models specified in Figure 4.1, a set of the six correlated frailties are required, one for each model. However, for identifiability and computational concerns, we impose some restrictions and simplifying assumptions. We initially propose two different formulations of the sets of models, and for ease of notation, we exclude baseline covariates X.

Model A using Time Dependent Covariates

For $z = 0$,

$$\lambda_{12}^0(t|\omega_{12i}^0) = \lambda_{12,0}^0(t)\exp(\kappa_{12}^0\omega_{12i}^0) \tag{4.2}$$

$$\lambda_{13}^0(t|\omega_{13i}^0) = \lambda_{13,0}^0(t)\exp(\kappa_{13}^0\omega_{13i}^0)$$

$$\lambda_{23}^0(t|T_{12i}(0),\omega_{23i}^0) = \lambda_{23,0}^0(t-T_{12i}(0))\exp(\kappa_{23}^0\omega_{23i}^0+\theta_{23}^0T_{12i}(0))I(t>T_{12i}(0))$$

Similarly for $z = 1$,

$$\lambda_{12}^1(t|\omega_{12i}^1) = \lambda_{12,0}^1(t)\exp(\kappa_{12}^1\omega_{12i}^1)$$

$$\lambda_{13}^1(t|\omega_{13i}^1) = \lambda_{13,0}^1(t)\exp(\kappa_{13}^1\omega_{13i}^1)$$

$$\lambda_{23}^1(t|T_{12i}(1),\omega_{23i}^1) = \lambda_{23,0}^1(t-T_{12i}(1))\exp(\kappa_{23}^1\omega_{23i}^1+\theta_{23}^1T_{12i}(1))I(t>T_{12i}(1))$$

where T_{12i} is the time that subject i moves into state S. Here we include θ_{23} in the λ_{23} model as the coefficient for our time dependent covariate T_{12}. The purpose is to capture the effect of this transition time, and the time that an individual experiences S may help to assess the strength of association between S and T. We model the transition using a clock reset for λ_{23} (ie the time scale is $t - T_{12}(z)$).

The restrictions and assumptions we will be considering are to make $\omega_{13i}^z = \omega_{23i}^z$ and to set some of the $\kappa_{jk}^z = 1$. If the κ parameters vary, they essentially influence how variable the frailty terms are. We will refer to κ as frailty coefficients. One rationale for assuming $\omega_{13i}^z = \omega_{23i}^z$ in this setting is that both are frailties that influence time to death from others causes in our motivating trial. For example, since our variable T is death from any cause, we may expect that some men will die of old age. It may be reasonable to expect that an individual may have their own propensity for experiencing death from other causes irrespective of whether or not S has occurred. Another

consideration is that by including the coefficient for our time-varying covariate, θ_{23}^{z}, the model captures the magnitude of the effect for the time it takes to experience the intermediate outcome S. This makes it more plausible that certain frailties are equal and conditional independence assumptions may be more likely. Lastly, the frailties capture heterogeneity on the individual level. There may still be heterogeneity on the population level for the variability in the hazard of going from baseline to T or from S to T which can be reflected in the baseline hazards. We explore these variations in later sections.

Model B using Multiple Frailties in Place of Time Dependent Covariates

We include an alternate option to incorporate the dependence between the different transitions such as a model that includes two frailty terms in the $S \to T$ transition

$$\lambda_{12}^{0}(t|\omega_{12i}^{0}) = \lambda_{12,0}^{0}(t)\exp(\kappa_{12}^{0}\omega_{12i}^{0}) \tag{4.3}$$

$$\lambda_{13}^{0}(t|\omega_{13i}^{0}) = \lambda_{13,0}^{0}(t)\exp(\kappa_{13}^{0}\omega_{13i}^{0})$$

$$\lambda_{23}^{0}(t|T_{12i}(0),\omega_{13i}^{*0},\omega_{12i}^{*0}) = \lambda_{23,0}^{0}(t-T_{12i}(0))\exp(\kappa_{12}^{*0}\omega_{12i}^{0}+\kappa_{13}^{*0}\omega_{13i}^{0})I(t>T_{12i}(0))$$

$$\lambda_{12}^{1}(t|\omega_{12i}^{1}) = \lambda_{12,0}^{1}(t)\exp(\kappa_{12}^{1}\omega_{12i}^{1})$$

$$\lambda_{13}^{1}(t|\omega_{13i}^{1}) = \lambda_{13,0}^{1}(t)\exp(\kappa_{13}^{1}\omega_{13i}^{1})$$

$$\lambda_{23}^{1}(t|T_{12i}(1),\omega_{13i}^{*1},\omega_{12i}^{*1}) = \lambda_{23,0}^{1}(t-T_{12i}(1))\exp(\kappa_{12}^{*1}\omega_{12i}^{1}+\kappa_{13}^{*1}\omega_{13i}^{1})I(t>T_{12i}(1))$$

The motivation of this model is an alternative way to capture the subject specific relationship between the different transitions via the κ_{12}^{*} and κ_{13}^{*} coefficients. This model does not include T_{12} as a time-varying covariate. When we assume $\omega_{23}^{z} = \omega_{13}^{z}$, the key difference between models A and

B would be how we parameterize the way in which the transition from baseline to the intermediate outcome and the time following that transition are related; these are linked using either a time varying covariate (in model A) or another frailty term (in model B). Again, the frailty coefficients κ can be thought of parameters that increase or decrease the magnitude of the effect of the frailties. We would not expect κ_{12}^{*z} and κ_{12}^{z} to be necessarily equal across the models given the different assumptions in each model.

Here we will first consider the six correlated frailties in model A.

$$\begin{pmatrix} \omega_{12i}^0 \\ \omega_{12i}^1 \\ \omega_{13i}^0 \\ \omega_{13i}^1 \\ \omega_{23i}^0 \\ \omega_{23i}^1 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_S & \rho_{00} & \rho_{01} & \rho_{S1} & \rho_{S2} \\ & 1 & \rho_{10} & \rho_{11} & \rho_{S3} & \rho_{S4} \\ & & 1 & \rho_T & \rho_{T1} & \rho_{T2} \\ & & & 1 & \rho_{T3} & \rho_{T4} \\ & & & & 1 & \rho_{ST} \\ & & & & & 1 \end{pmatrix} \right)$$

While this model has a very general form, it may not be necessary or even desirable to consider this level of generality. We will be focusing on special cases of this general model, which we think are appropriate for the setting of surrogacy assessment.

To reduce the number of frailties to estimate to four in model A, we assume that both transitions into T have the same frailty ($\omega_{13}^z = \omega_{23}^z$) since they are both relevant for time to the terminal event. As discussed above, since the terminal event is death from any cause, it seems justifiable to assume that conditional on all other terms in the model, frailties toward death from any cause would be the same on the individual level with or without spreading of the cancer. This assumption will be useful for estimation since T_{23i} is not defined for all individuals. With this assumption, our

transition models from S to T in model A can be written

$$\lambda^0_{23}(t|T_{12i}(0), \omega^0_{13i}) = \lambda^0_{23,0}(t - T_{12i}(0)) \exp(\kappa^0_{23}\omega^0_{13i} + \theta^0_{23}T_{12i}(0))I(t > T_{12i}(0))$$

$$\lambda^1_{23}(t|T_{12i}(1), \omega^1_{13i}) = \lambda^1_{23,0}(t - T_{12i}(1)) \exp(\kappa^1_{23}\omega^1_{13i} + \theta^1_{23}T_{12i}(1))I(t > T_{12i}(1))$$

Ideally, we would like to allow κ^z_{23} to take on different values from κ^z_{13} to accommodate different amounts of dependence between the transitions. For both models A and B we consider the joint distribution

$$\begin{pmatrix} \omega^0_{12i} \\ \omega^1_{12i} \\ \omega^0_{13i} \\ \omega^1_{13i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_S & \rho_{00} & \rho_{01} \\ & 1 & \rho_{10} & \rho_{11} \\ & & 1 & \rho_T \\ & & & 1 \end{pmatrix}\right)$$

In most of the work presented in this chapter, we will also assume $\omega^z_{12i} \perp \omega^z_{13i}$ (the frailties for an individual are independent across states), meaning $\rho_{00} = \rho_{01} = \rho_{11} = \rho_{10} = 0$. We thus assume

$$\begin{pmatrix} \omega^0_{12i} \\ \omega^1_{12i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_S \\ \rho_S & 1 \end{pmatrix}\right) \qquad \begin{pmatrix} \omega^0_{13i} \\ \omega^1_{13i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_T \\ \rho_T & 1 \end{pmatrix}\right) \tag{4.4}$$

This type of independence assumption may aid in estimation. We could instead impose a strong assumption in the opposite direction that $\omega^z_{12} = \omega^z_{13}$. Another possible assumption would be to consider a single frailty for each arm, ie $\omega^0_{12i} = \omega^0_{13i} = \omega^0_{23i}$ and $\omega^1_{12i} = \omega^1_{13i} = \omega^1_{23i}$. The motivation for this comes from considering the frailty as representing an omitted covariate. We do not further pursue this assumption.

4.2.2 Identifiability and Sensitivity Analysis

Within an observed arm, we can evaluate the sensitivity of the assumed models. For example, we can vary which frailties are assumed to be independent or equal, alter which values of κ_{jk}^{z} are set to 1, change the baseline hazard from a Weibull distribution to piecewise exponential or something more flexible, assess different effects of covariates in the transitions, and modify the time-reset parameterization. The parameters ρ_S and ρ_T are not identifiable, so they will be fixed at preset values. Based on biological considerations under the counterfactual framework, we may not expect these correlation parameters to be negative or exactly equal to 1. We provide a tool for assessing the sensitivity of these values and commentary on the feasibility and identifiability of estimating these models with and without these assumptions in later sections.

4.3 Likelihood and Estimation

4.3.1 Likelihood Contributions

We consider a randomized clinical trial of n subjects for a binary treatment Z. For generality, let n_z denote the number of subjects in treatment arm $Z = z$ (and we may assume that $n/2$ subjects are in treatment group $z = 1$ and $n/2$ are in treatment group $z = 0$ since the treatment assignment is randomized and under the control of the investigator). Let $\{S_i, \delta_{Si}, T_i, \delta_{Ti}, X_i, Z_i\}$ be the observed data for subject i for $i = 1, ..., n$. We will also consider a random or administrative censoring time C_i. S_i denotes the time to transition to state S, T_i denotes the time that the terminal event T occurs, and δ_T and δ_S denote the censoring indicators for $T(z)$ and $S(z)$ being observed. Then $\delta_{Ti} = 1$ when $T_i < C_i$ and $\delta_{Si} = 1$ when $S_i < C_i$ and $S_i < T_i$.

We can also conceptualize the data in terms of the random variables in Figure 4.1. Based on gap

times between states T^z_{jk}, the data can also be represented as $\{T_{12i}, T_{13i}, T_{23i}, \delta_{Si}, \delta_{Ti}, X_i, Z_i\}$. In the illness-death formulation, there are four possible combinations of observable δ_{Si} and δ_{Ti}. We assume that when neither event is observed, meaning $\delta_{Si} = \delta_{Ti} = 0$, then $T_{12i}(z)$ and $T_{13i}(z)$ take on the same value as being censored at C_i. Notably, T_{23i} is not defined when S_i is not observed. Consider when T is observed before S, meaning $\delta_{Ti} = 1, \delta_{Si} = 0$. Then the observed data related to S_i for individual i is equal to $\{T_{13i}, \delta_{Si} = 0\}$, and observed T_i is based on $\{T_{13i}, \delta_{Ti} = 1\}$, while T_{23i} is not defined. Now consider when only S is observed, meaning $\delta_{Ti} = 0, \delta_{Si} = 1$. Then the observed data for individual i is S_i based on $\{T_{12i}, \delta_{Si} = 1\}$. Assuming T is not observed after, the value T_i takes on is censored at $\{C_i, \delta_{Ti} = 0\}$. If both S and T are observed with $\delta_{Ti} = \delta_{Si} = 1$, then S_i is based on $\{T_{12i}, \delta_{Si} = 1\}$, and T_i is based on $\{T_{12i} + T_{23i}, \delta_{Ti} = 1\}$. We provide the likelihood under these scenarios next.

We assume that each hazard in Figure 4.1 follows a Weibull distribution, so we have $T_{12}(z) \sim Weibull(\alpha^z_{12}, \gamma^z_{12}), T_{13}(z) \sim Weibull(\alpha^z_{13}, \gamma^z_{13})$, and $T_{23}(z) \sim Weibull(\alpha^z_{23}, \gamma^z_{23})$ for shape parameters α^z_{jk} and scale parameters γ^z_{jk}. The scale and shape parameters must be positive: $\gamma^z_{jk} > 0, \alpha^z_{jk} > 0$. We parameterize the cumulative baseline hazard function as $\Lambda^z_{jk0}(t) = \gamma^z_{jk} t^{\alpha^z_{jk}} = \int_0^t \lambda^z_{jk0}(u)du$ for a given Weibull model, where $\lambda^z_{jk0}(t) = \gamma^z_{jk}\alpha^z_{jk} t^{\alpha^z_{jk}-1}$ and $\lambda^z_{jk}(t) = \lambda^z_{jk0}(t)\exp(\kappa^z_{jk}\omega^z_{jk})$ for jk = 12 or 13. The model for λ^z_{23} is more complex than this and depends on whether model A or B is assumed.

We will consider the likelihood of the observed data for each arm separately. For ease of notation, we will drop the superscript in this section as the derivations apply to both treatment arms. An alternative approach is to base estimation on the complete data likelihood, where that likelihood is derived using the random variables in Figure 4.1 with both sets of counterfactual outcomes under the two treatment arms $T_{12}(0), T_{12}(1), T_{13}(0), T_{13}(1), T_{23}(0), T_{23}(1)$. However, an important distinction is that this approach would jointly model the outcomes and involve all

elements ρ of the correlation matrix in equation 4.4. Using this specification as an alternative form of the likelihood, an imputation scheme could be proposed to fill in all missing outcomes. Any relation between the potential outcomes across treatment arms for an individual in the complete data likelihood is not identified. Based on previous exploration of methods that use either the observed or the complete data likelihood in this dissertation, using this complete data likelihood is not necessary. Here we will only focus on the observed data likelihood during estimation. Any counterfactual quantities needed for calculation of the CEP curve will be described separately in Section 4.4. We note that $\{T_{23i}, \omega_{23i}\}$ are not defined when $\delta_{Si} = 0$ and do not contribute to the likelihood, which is the case for either the complete data or observed data likelihood.

The likelihood contributions can be written similarly to work done by Conlon et al. (2014b). For those who had not experienced S, we are in the setting where $\delta_{Si} = 0$, and T_{23i} is not defined:

$\lambda_{13}(T_{13})^{\delta_T} \exp(-\int_0^{T_{13}} \lambda_{13}(u)du - \int_0^{T_{13}} \lambda_{12}(u)du)$

For those who experience S, and are either dead or alive, $\delta_{Si} = 1$, and T_{23i} is defined. δ_{Ti} may be equal to either 0 or 1 depending on if the terminal event is observed:

$\lambda_{12}(T_{12}) \exp(-\int_0^{T_{12}} \lambda_{12}(u)du - \int_0^{T_{12}} \lambda_{13}(u)du)\lambda_{23}(T_{23}|T_{12})^{\delta_T} \exp(-\int_0^{t_{23}} \lambda_{23}(u|T_{12})du)$

4.3.2 Bayesian Estimation

To facilitate estimation, we will take a Bayesian approach using Markov Chain Monte Carlo (MCMC). We use prior distributions that are similar to those suggested in Gao et al. (2012) and Sahu et al. (1997). Regression coefficients are assumed to have a diffuse normal prior distribution (Sahu et al. 1997). We assume a Gamma(p_1, p_2) prior for the scale parameters γ_{jk} of the Weibull distribution, and we also assume a Gamma(p_3, p_4) prior for the shape parameters α_{jk} with hyperparameters $p_1 = p_2 = p_3 = p_4 = 0.1$.

Any parameters that do not have a closed-form posterior distribution $(\alpha^z_{jk}, \gamma^z_{jk}, \omega^z_{jk}, \theta^z_{23}, \kappa^z_{23})$

are drawn using a Metropolis-Hastings step (Robert and Casella, 2004). At each iteration of the MCMC, proposed draws of the parameters are taken from a Gaussian proposal distribution π with mean equal to the previous accepted draw. For a general parameter β and iteration p of the MCMC, we draw a proposed value of $\beta^{'} \sim N(\beta^{p-1}, \sigma^2)$ based on using the previous iteration β^{p-1}. The acceptance ratio is calculated as $\frac{P(\beta^{'})}{P(\beta^{p-1})} \times \frac{\pi(\beta^{'})}{\pi(\beta^{p-1})}$ where $P(\beta)$ represents the posterior distribution of β and π represents the proposal density. For a general Gaussian density, $g(\beta^{'}|\beta^{p-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-1/2\sigma^2)(\beta^{'} - \beta^{p-1})^2$ and $g(\beta^{p-1}|\beta^{'}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-1/2\sigma^2)(\beta^{p-1} - \beta^{'})^2$. Based on our proposal distribution, the exponential terms in the ratio of Gaussian densities will cancel, so when we calculate the ratio $P(\beta^{'})/P(\beta^{p-1})$, the proposed draw $\beta^{'}$ is accepted with the simplified probability $min(1, \frac{P(\beta^{'})}{P(\beta^{p-1})})$. The variance of the proposal distribution σ^2 is tuned to obtain convergence of parameter draws and target a reasonable acceptance rate (Gelman et al. 1996).

The estimated frailties are also drawn using a Metropolis-Hastings step with a Gaussian proposal distribution and a Gaussian prior with mean zero and standard deviation equal to 1. Each proposed frailty term for an individual has its own acceptance ratio. For $i = 1, ..., \frac{n}{2}$, we obtain draws of $\omega^0_{12i}, \omega^0_{13i}$, and for $i = \frac{n}{2} + 1, ..., n$, we obtain draws of $\omega^1_{12i}, \omega^1_{13i}$ using the posterior distribution.

The likelihood contributions for L for each parameter can be found in Appendix P. Based on the given likelihood components and prior distributions π^*, the posterior P for a given $Z = z$ is the product over individuals i:

$$\prod_i (L_i(T_{13i}(z), T_{23i}(z), T_{12i}(z), \delta_{Si}, \delta_{Ti}, \omega^z_{12i}, \omega^z_{13i}, \omega^z_{23i}, \beta^z_{12}, \gamma^z_{12}, \alpha^z_{12}, \beta^z_{13}, \gamma^z_{13}, \alpha^z_{13}, \beta^z_{23}, \gamma^z_{23}, \alpha^z_{23}, \theta^z_{23}, \kappa^z_{12}, \kappa^z_{13}, \kappa^z_{23}) \times$$

$$\pi^*(\omega_{12i})\pi^*(\omega_{13i})\pi^*(\omega_{23i})\pi^*(\beta_{12})\pi^*(\gamma_{12})\pi^*(\alpha_{12})\pi^*(\beta_{13})\pi^*(\gamma_{13})\pi^*(\alpha_{13})\pi^*(\beta_{23})\pi^*(\gamma_{23})\pi^*(\alpha_{23})\pi^*(\theta_{23})\pi^*(\kappa_{12})\pi^*(\kappa_{13})\pi^*(\kappa_{23}))$$

Visually, we can see the hierarchy of parameters across different treatments and transitions and how the terms are related in Figure 4.2.

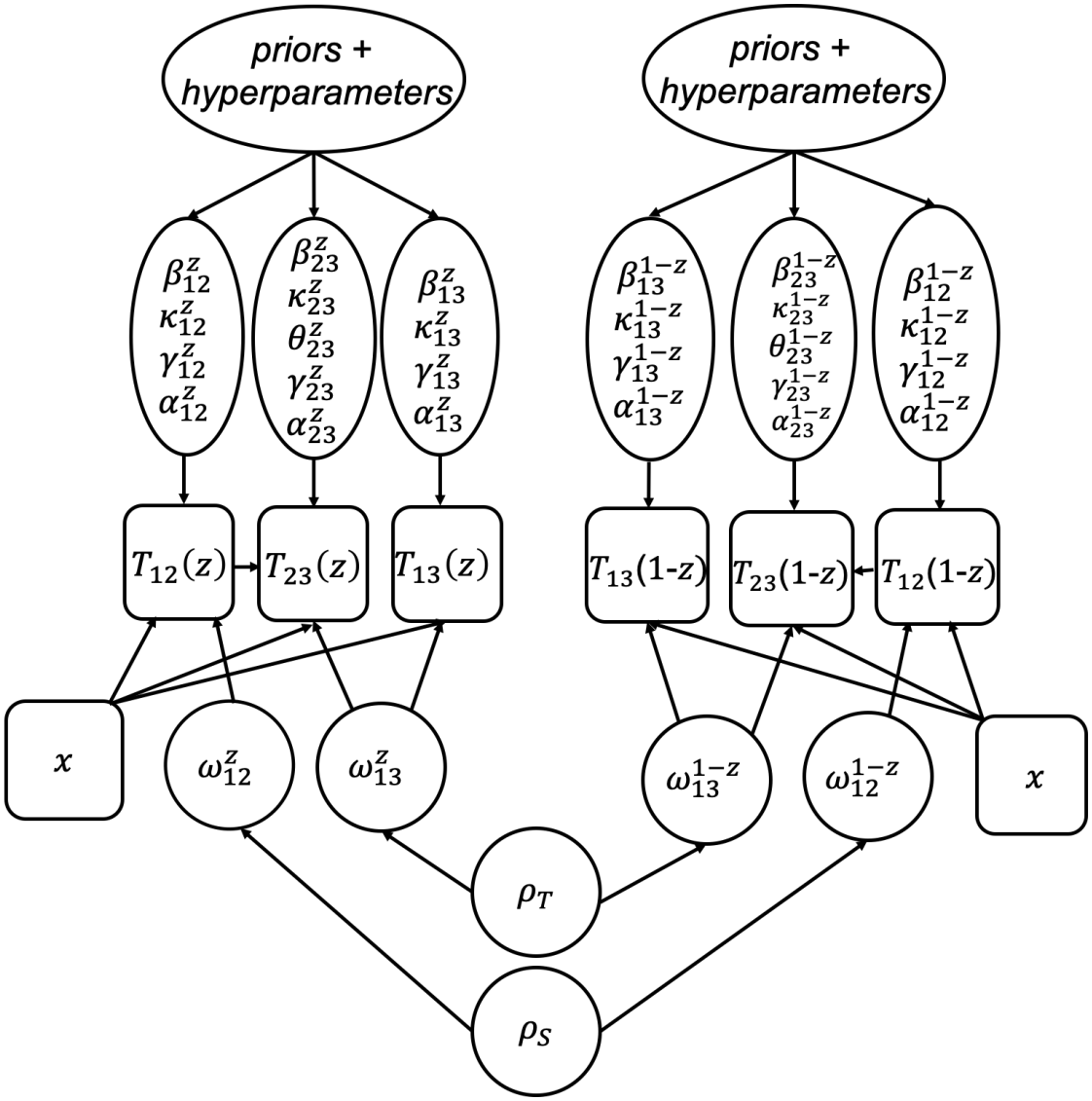


Figure 4.2: This diagram demonstrates the relationships between the parameters and data in the proposed model (model A assuming that $\omega_{13}^z = \omega_{23}^z$).

4.4 CEP Quantities

We develop a method for validating a surrogate endpoint using the principal stratification framework (Frangakis and Rubin, 2002). The goal of this validation procedure is to develop causal quantities that rigorously determine if a time-to-event S is a valid surrogate for use in a future trial

in place of T. In a non-survival setting, Gilbert and Hudgens (2008) define a principal surrogate endpoint for a binary T based on the comparison of the quantities $risk_{(1)}(s_1, s_0) \equiv P(T(1) = 1|S(1) = s_1, S(0) = s_0)$ and $risk_{(0)}(s_1, s_0) \equiv P(T(0) = 1|S(1) = s_1, S(0) = s_0)$. The condition that these must be equal for all $s_1 = s_0$ is known as average causal necessity. Average causal sufficiency is defined as $risk_{(1)}(s_1, s_0) \neq risk_{(0)}(s_1, s_0)$ for all $|s_1 - s_0| > C$ for some non-negative constant C. They define the causal effect of the treatment on the true endpoints as $h(P(T(1) = 1), P(T(0) = 1))$ for some $h(,)$ contrast function that satisfies $h(x, y) = 0$ if and only if $x = y$. The CEP surface is therefore equal to $h(risk_{(1)}, risk_{(0)})$ over values of $s = (s_1, s_0)$. A specific case of this is the CEP plot of $\Delta T = E(T(1) - T(0)|S(1) - S(0) = s)$ over values of $\Delta S = S(1) - S(0) = s$ when S and T are continuous. Based on these criteria, an ideal CEP plot for a valid surrogate will go through the origin and have a positive slope. We generalize this by defining new contrasts, ΔT_i and ΔS_i for each subject in this time-to-event setting, forming a scatterplot of $(\Delta S_i, \Delta T_i)$, and assessing whether a line through the points on this scatterplot goes through the origin and has a positive slope. For ΔT we will use $P(T(1) > \tau_T) - P(T(0) > \tau_T)$ evaluated at time τ_T. For ΔS we will use $\log\left(\frac{\Lambda^0_{12}(\tau_S)}{\Lambda^1_{12}(\tau_S)}\right)$ that depends on some time τ_S. τ_S and τ_T must be chosen at meaningful or sensible times. For example, for a surrogate to be useful, it is likely that $\tau_S < \tau_T$. These times should be chosen to be evaluated after a sufficient number of events have occurred in order to make sensible decisions about the surrogate. It is also possible to use the same estimands for both ΔS and ΔT. Here we have chosen this ΔT as a more interpretable quantity that can be calculated regardless of whether S has occurred. It is also possible to express ΔS as a probability using a simple transformation of the cumulative hazards if desired.

While counterfactual draws of the frailties are not needed for the estimation procedure, they are needed for the proposed CEP formulation. As the correlations between the observed and counterfactual outcomes are non-identified, we fix ρ_S, ρ_T from the distributions in equation 4.4 to draw

the counterfactual frailty terms. We use correlations of 0.5 as a starting point since it is a mid-point between perfect and no correlation, and we also vary ρ_S and ρ_T for sensitivity analysis. We use the prior distribution and fixed ρ_S, ρ_T to obtain draws of the ω estimates in the counterfactual arm from the appropriate conditional normal distributions, such as $\omega_{12}^z|\omega_{12}^{1-z} \sim N(0+\rho_S(\omega_{12}^{1-z}), 1-\rho_S^2)$ and similarly $\omega_{13}^z|\omega_{13}^{1-z} \sim N(0+\rho_T(\omega_{13}^{1-z}), 1-\rho_T^2)$. We repeat the process for the other treatment arm to obtain sets of counterfactual frailties for each individual.

Each individual has a set of subject-specific hazards that will be used in a CEP plot. Let $\Delta S_i = \log \frac{\Lambda_{12}^0(\tau_S|\omega_{12i}^0, x_i)}{\Lambda_{12}^1(\tau_S|\omega_{12i}^1, x_i)}$ be on the x-axis of the plot where $\Lambda_{12}^0(\tau_S|\omega_{12}^0, x) = \int_0^{\tau_S} \lambda_{12}^0(t|\omega_{12}^0, x)dt$ and $\Lambda_{12}^1(\tau_S|\omega_{12}^1, x) = \int_0^{\tau_S} \lambda_{12}^1(t|\omega_{12}^1, x)dt$. For the y-axis, consider $\Delta T_i = P(T_i(1) > \tau_T|\omega_{12i}^1, \omega_{13i}^1, \omega_{23i}^1) - P(T_i(0) > \tau_T|\omega_{12i}^0, \omega_{13i}^0, \omega_{23i}^0)$ based on the frailties in model A. For example, using model A, $\Delta S_i = \log \frac{\Lambda_{12,0}^0(t)\exp(\kappa_{12}^0\omega_{12i}^0)}{\Lambda_{12,0}^1(t)\exp(\kappa_{12}^1\omega_{12i}^1)}$.

Overall survival at time τ can be decomposed into components based on P(do not experience S or T) + P(experience S but not T). More formally, this framework is similar to the likelihood for a joint illness-death model proposed in Suresh et al. (2017) and for illness-death with a cure fraction proposed by Conlon et al. (2014b) and Beesley et al. (2019). In formal notation, we are interested in the quantities

$$P(T(0) > \tau_T) = P(T(0) > \tau_T, S(0) > \tau_T) + P(T(0) > \tau_T, S(0) < \tau_T) =$$

$$P(T(0) > \tau_T|S(0) > \tau_T)P(S(0) > \tau_T) + P(T(0) > \tau_T|S(0) < \tau_T)P(S(0) < \tau_T)$$

and similarly for

$$P(T(1) > \tau_T) = P(T(1) > \tau_T, S(1) > \tau_T) + P(T(1) > \tau_T, S(1) < \tau_T) =$$

$$P(T(1) > \tau_T | S(1) > \tau_T) P(S(1) > \tau_T) + P(T(1) > \tau_T | S(1) < \tau_T) P(S(1)) < \tau_T)$$

These quantities can be written in terms of parameters

$$\exp(-\int_0^{\tau_T} \lambda_{12}(u)du - \int_0^{\tau_T} \lambda_{13}(u)du) + \int_0^{\tau_T} \exp(-\int_0^u \lambda_{12}(v)dv - \int_0^u \lambda_{13}(v)dv)\lambda_{12}(u) \exp(-\int_0^{\tau_T - u} \lambda_{23}(v|u)dv)du$$

$$= \exp(-\Lambda_{12}(\tau_T) - \Lambda_{13}(\tau_T)) + \int_0^{\tau_T} \exp(-\Lambda_{12}(u) - \Lambda_{13}(u))\lambda_{12}(u) \exp(-\int_0^{\tau_T - u} \lambda_{23}(v|u)dv)du$$

Based on the draws of all model parameters for a given iteration of the MCMC, we estimate observed and counterfactual hazards for each individual. After calculating ΔT_i and ΔS_i conditional on the set of ω_i, we create a scatterplot of ΔT_i vs. ΔS_i and draw a loess or linear curve through the points for a single iteration of the algorithm. Our γ_0 and γ_1 summary quantities are equal to the intercept and slope of this line (and these quantities may need to be redefined for a loess curve). This process is repeated for the next set of random draws of model parameters and frailties for all individuals. These quantities are then averaged over the iterations of the MCMC after a burn-in period.

4.4.1 Valid Surrogates under an Illness-Death CEP Curve

As our CEP curve is a fairly complex function of these quantities, we empirically investigate what combination of illness-death models, meaning relationship between S and T, leads to CEP plots that align with an intuitive notion of whether S is a good surrogate for T. We consider the eight scenarios that may exist based on which transitions have treatment effects (defined as whether or not the counterfactual hazards are equal) in Table 4.1.

In addition to which transitions have hazards that are moderated by treatment, each combination can be crossed with the effect of θ_{23} and κ_{23} being zero vs. nonzero in a factorial de-

sign. We characterize the CEP curves under these scenarios using true generating parameter values to calculate ΔT and ΔS. In Appendix N, we show scatterplots of ΔS_i vs. ΔT_i for simulated data, for which the values of the frailties are known. An Rshiny app is also available at https://emilyroberts.shinyapps.io/id_cep_parameters/ that allows users to characterize the CEP curve for different parameter values. We also allow for the user to vary which independence or equivalence assumptions are made about the frailty terms and the corresponding impact on the CEP curve.

Based on these settings, we suggest which data scenarios should correspond to a decision that the intermediate outcome is in fact a valid surrogate. We identify that for a perfect surrogate, the paths that treatment effects should exist are through the baseline to intermediate outcome transition only (ie $\lambda_{12}^0 \neq \lambda_{12}^1$). In the null case, Scenario 1, and this ideal case Scenario 2, the estimated slope is positive and the intercept is equal to 0. This is consistent with our consideration of the more flexible Prentice Criteria, which also suggest that hazards from baseline to S should be non-equal ($\lambda_{12}^0 \neq \lambda_{12}^1$) and the hazards from baseline to T should be equal ($\lambda_{13}^0 = \lambda_{13}^1$) across treatment arms. We can also examine the marginal effects on S and T for these scenarios. For scenario 1, they are both zero as expected. For scenario 2, the marginal effect on T is rather small under these parameter values. Further, while we anticipated differences between perfect, partial, and non-surrogates would be easily apparent, the slope does not drastically change between the different scenarios. Under the particular parameters we investigated, the slope may be positive for all of the scenarios. We did observe that Scenarios 3-8 (denoted as partial and non-surrogates) produced CEP curves that did not go through the origin and therefore were invalid.

Possible explanations for the small differences in slope values across the scenarios include that the y-axis will always be constrained between -1 and 1 since it represents a difference in two probabilities. This quantity ΔT_i on the y-axis is a relatively complex function of multiple

model parameters that may not change drastically based on relatively small changes in the baseline hazards. We do see that incorporating non-zero values of θ_{23}^z does change the slope and intercept of the CEP curve in Figures 14.4 in the appendix. In other settings, we also find that the relative magnitude of the baseline hazards for $T_{12}(z), T_{13}(z)$, and $T_{23}(z)$ for a given treatment arm also influences the slope and intercept of a CEP curve. Largely, slightly changing the values of ρ in the correlation matrix of the frailty terms does not have a major impact on the CEP slope and intercepts, though other settings in the online app demonstrate specific settings where these correlations may be more consequential.

4.5 Simulation Study

4.5.1 Simulation Set-up

Here we start with a simulation setting where we assume each baseline hazard follows a Weibull distribution where shape parameters for the baseline hazards and frailty coefficients are equal to 1. We conduct a simulation with 100 replicated datasets and $n = 600$. Data are generated under simple settings that follow the θ parameterization shown in model A. Survival times are simulated based on a Weibull baseline hazard specification (Austin, 2012). We generate the frailties to have mean 0 and a standard deviation of 1. We will describe our assumptions about the frailties, where we assume and generate them such that $\omega_{13}^z = \omega_{23}^z$ in most settings.

We conduct the estimation procedure described in section 4.4. Initial estimates of the frailties may be calculated using the `frailtypack` or `frailtyEM` packages in `R` (R Core Team; Rondeau and Gonzalez, 2005; Balan and Putter, 2019). Parameter estimates are each drawn from the proposal distribution individually. Under the parameterization in model A, θ_{23}^z is drawn from a proposal distribution with a mean based on the estimated coefficient from a haz-

ard model fit using observed data regressing time to T on time to S, among those who experience S. By doing this, θ_{23}^1 and θ_{23}^0 have unique starting values. The draws are accepted in blocks for the Metropolis-Hastings step. The blocks are divided into treatment arm transitions, and the parameters within a block are jointly accepted or rejected. For model A, we have blocks $\omega_{12i}^0; \{\gamma_{12}^0, \alpha_{12}^0\}; \omega_{13i}^0; \{\gamma_{13}^0, \alpha_{13}^0\}; \{\gamma_{23}^0, \alpha_{23}^0, \theta_{23}^0, \kappa_{23}^0\}; \omega_{12i}^1; \{\gamma_{12}^1, \alpha_{12}^1\}; \omega_{13i}^1; \{\gamma_{13}^1, \alpha_{13}^1\}; \{\gamma_{23}^1, \alpha_{23}^1, \theta_{23}^1, \kappa_{23}^1\}$ when all of the model parameters are being estimated. The proposal distributions have standard deviation $\sigma = 0.1$.

The true values of the parameters are shown in the simulation results in the first row of each table of results. Current simulation studies are shown for simple settings where true values are $\kappa_{23}^z = \kappa_{12}^z = \kappa_{13}^z = 1$. In all cases we fix the shape parameters $\alpha_{jk}^z = 1$ during estimation (essentially assuming an exponential distribution). Based on identifiability of the baseline hazard, frailties, and coefficients associated with the frailties, we consider two options: fix all scale parameters γ_{jk}^z and estimate the frailty coefficients κ_{23}^z or fix κ_{23}^z and estimate the scale parameters. In all explored simulations, we assume that $\kappa_{12}^z = \kappa_{13}^z$ equals the true value 1 for identifiability of the models. In our main set of simulation studies, we generate and assume during estimation that all $\kappa_{jk}^z = 1$, and estimate the scale parameters γ_{jk}^z and θ_{23}^z parameters.

We conduct simulation studies from the eight possible scenarios, highlighting Scenario 1 with no treatment effects, Scenario 2 where there is a treatment effect only on S (a perfect surrogate), and scenarios 3-8 where treatment effects exist such that we do not expect S to be a surrogate. We generate treatment effects by differing the scale parameters between arms, meaning $\gamma_{jk}^1 \neq \gamma_{jk}^0$. We also conduct some sensitivity analyses by varying the assumptions that $\omega_{12}^z \perp \omega_{13}^z$ and $\omega_{13}^z = \omega_{23}^z$. In these cases, we assume either that $\omega_{12}^z \perp \omega_{13}^z \perp \omega_{23}^z$ or that all three frailties are correlated within a given counterfactual treatment arm. In these settings, we estimate the set $\omega_{12i}^z, \omega_{13i}^z, \omega_{23i}^z$ for each individual. T_{23i}^z and corresponding ω_{23i} does not exist for any individual

that does not experience the intermediate event. In this case, ω_{23} is drawn directly from the prior or its conditional multivariate normal distribution using our most general model formulation with six frailties and a fixed covariance matrix. For example, we assume $\rho_{T1} = \rho_{T4} = 0.95$ and $\rho_{T3} = \rho_{T2} = \rho_{ST} = \rho_T = 0.5$. In these cases we set $\tau_S = 1$ and $\tau_T = 2$.

4.5.2 Simulation Results

In this section, we show results of the estimated model parameters as well as validation quantities, the intercept γ_0, and slope γ_1. The estimation of the γ_0 and γ_1 quantities are calculated from fitting a linear best fit line through the CEP cloud at each iteration and reporting the posterior mean of these quantities for each simulated dataset. Parameter estimates are based on the posterior means and corresponding measures of variability; the average estimated standard error (SE) and the standard deviation (SD) of the posterior means are shown for the model parameters. We run the simulations for 5,000 iterations with 500 burn in draws. In addition to trace plots of the parameter draws, we assess the empirical mean and standard deviation of the estimated frailty terms over the iterations.

As mentioned in Section 4.5.1, we consider one strategy to fix all scale parameters and estimate κ^z_{23} or fix κ^z_{23} and estimate the scale parameters. The estimates in Tables 17.1 and 17.2 in Appendix Q show a subset of these results. Since we are estimating κ_{23} in these models, we fix the value of γ^z_{jk} to the true value. This assumption is for illustrative purposes of the model as this is restrictive by essentially assuming the treatment effects are known. Exploration of different settings suggests it may become difficult to estimate κ^z_{23} well when γ^z_{23} is not fixed due to identifiability issues.

In the main set of simulations in Tables 4.2 and 4.3, we fix κ^z_{23} at its true value and estimate the scale parameters. In this case, $\gamma^z_{12}, \gamma^z_{13}, \gamma^z_{23}$, and θ_{23} are estimated well when initial estimates are reasonable. These identified model parameters seem to converge well based on the assumptions we have made in these simulations. The distribution of the estimated frailty terms can deviate from

the generating distribution with mean zero and fixed variance. While our method involves prior and proposal distributions for the frailties, we are not directly enforcing any assumptions about the mean or variability of the frailty parameters during the estimation algorithm. The shape of the likelihood for frailty terms, particularly ω_{12}^z terms for individuals with $\Delta_{Si} = 0$, seems to be fairly flat, so the draws move around considerably during the algorithm. In these considered simulations, the credible intervals around γ_1, γ_0 are somewhat wide for all scenarios. Since an ideal surrogate will have values $\gamma_0 = 0$ and $\gamma_1 > 0$, the uncertainty can make it difficult to determine the value of the surrogate. This may incorrectly lead us to draw the same conclusions about the surrogate under all scenarios.

In Figure 4.3 we show the CEP curve conditional on estimated frailties for one dataset from these studies under Scenario 2. Each point is the posterior mean of (ΔS_i, ΔT_i) across MCMC iterations. The posterior values of the slope and intercept are shown, which convey the amount of variability based on the posterior coordinates of (ΔS_i, ΔT_i) for each individual i. We see that the estimated slope and intercept correctly meet our criteria of a valid surrogate under our proposed set of model assumptions. Though there is substantial variability in the estimates of γ_0 and γ_1, the respective posterior mean and credible intervals are -0.044 (-0.108, 0.020) and 0.087 (0.066, 0.108) for this dataset.

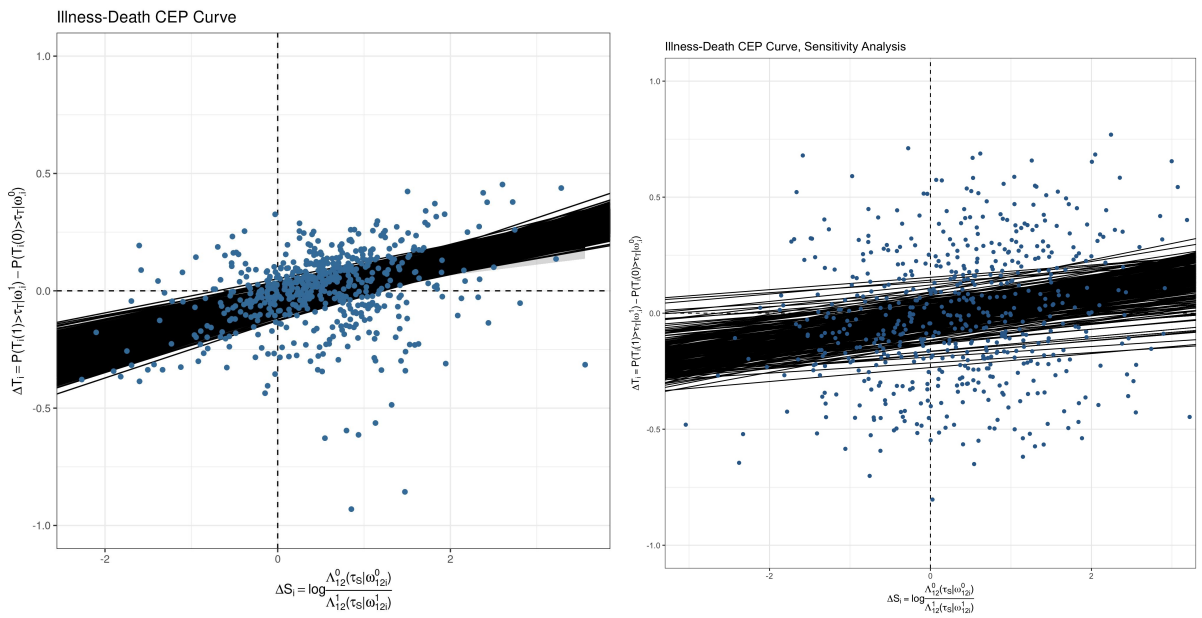


Figure 4.3: Example of an estimated CEP curve, conditional on frailties, for a single simulated dataset under Scenario 2. In this case, we assume all values of κ_{jk}^z are fixed at 1 and estimate the scale parameters in the Weibull distribution. On the left hand side, we assume that $\omega_{13}^z = \omega_{23}^z$ and $\omega_{13}^z \perp \omega_{12}^z$. On the right hand side, we check the sensitivity of these assumptions and allow all six counterfactual frailties for an individual to be unequal but correlated.

In our sensitivity analyses about the assumptions on the frailty terms, shown in Table 4.4, we see some sensitivity to the assumptions being made, such as increased variability in the subject-specific points. We show the results of one dataset under our sensitivity analyses in Figure 4.3. How these factors influence the CEP curves should be investigated under trial specific contexts.

4.6 Data Example

Our motivating clinical study is a phase III, randomized trial for men with prostate cancer, NRG/RTOG 9601 (Shipley et al., 2017). The trial features 760 men with recurrently or persistently elevated prostate-specific antigen (PSA) prostate cancer. The two treatments being compared are post-prostatectomy radiation therapy with or without antiandrogen therapy. There are 384 and 376

men in each treatment arm. The two survival endpoints of interest are time to distant metastasis, defined as radiographic evidence of metastatic cancer, and overall survival (OS). Notably, composite endpoints such as metastasis-free survival (MFS) are often evaluated. However, within our illness-death framework we consider time to distant metastasis and time to death separately. It has been previously established by The Intermediate Clinical Endpoints in Cancer of the Prostate (ICECaP) that MFS is a valid surrogate for OS in the setting of the initial treatment for localized prostate cancer (Xi et al., 2017). Others have evaluated if MFS is a valid surrogate when assessing the impact of antiandrogen therapy in recurrent prostate cancer following post-prostatectomy salvage radiation therapy (Jackson et al., 2020). Covariates in the dataset are also available, including PSA values at the time of randomization, Gleason score, and age in grouped categories.

We show in Figure 4.4 the Kaplan Meier curves for the intermediate and true outcomes without considering the semi-competing risk as well as the curve for the transition from S to T for those who experienced distant metastasis. S may be censored because it was not observed during the study period or because the terminal event T occurred first. We also present the cumulative incidence curve for S considering T as a semi-competing risk. The cumulative incidence estimates are based on the non-parametric Aalen-Johansen estimate of the cumulative incidence function from the `mstate` `R` package (Putter, 2011). The set of curves in Figure 4.5 are stratified into the two treatment groups. The purpose of this figure is to overcome the inability to interpret the curves for time to S in Figure 4.4 as probabilities due to the semi-competing risk structure.

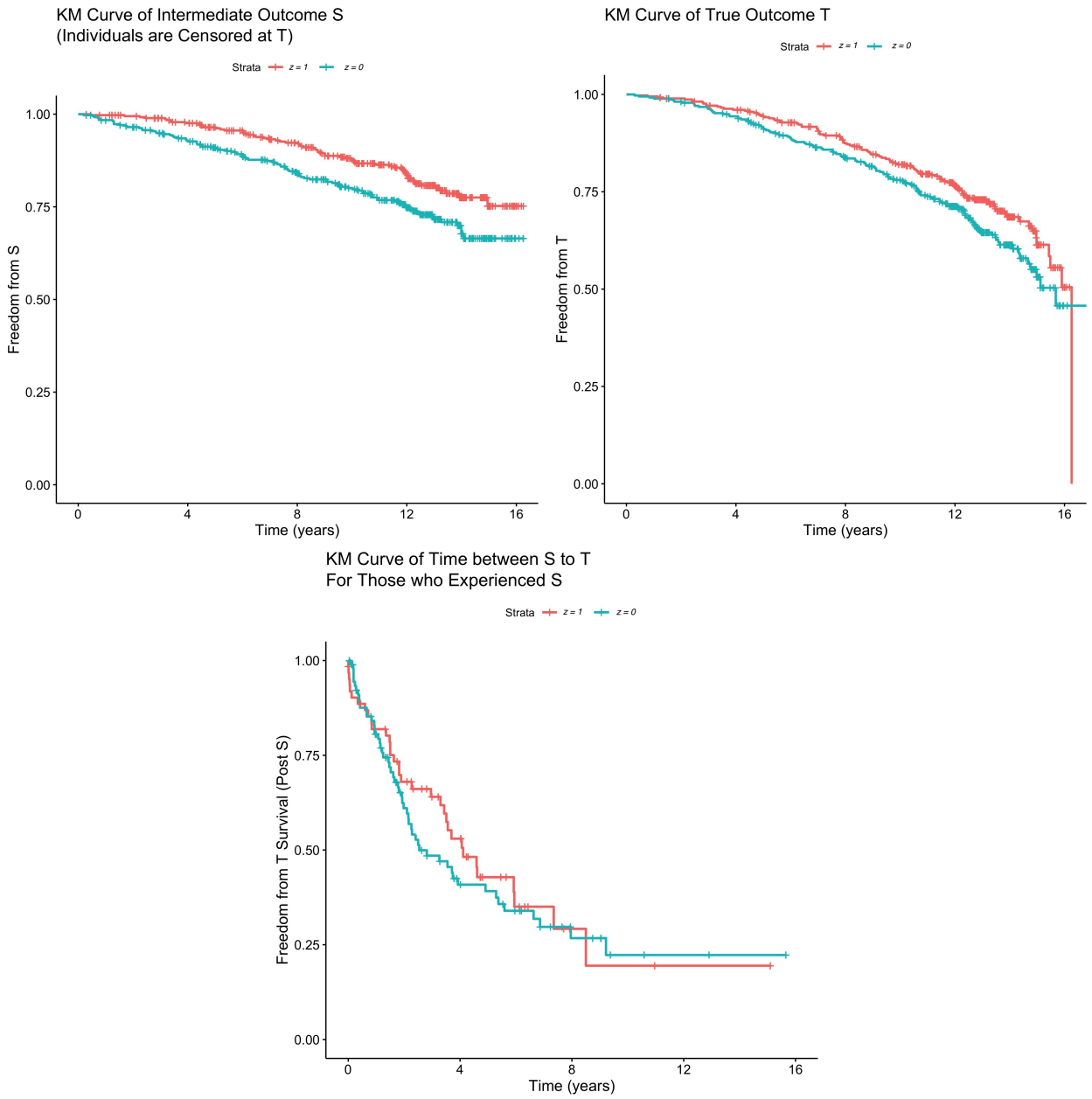


Figure 4.4: Kaplan Meier curves for the intermediate and true outcome demonstrating significant treatment effects for the prostate cancer trial. We also show the Kaplan Meier curve for the transition from S to T among those who experienced S.

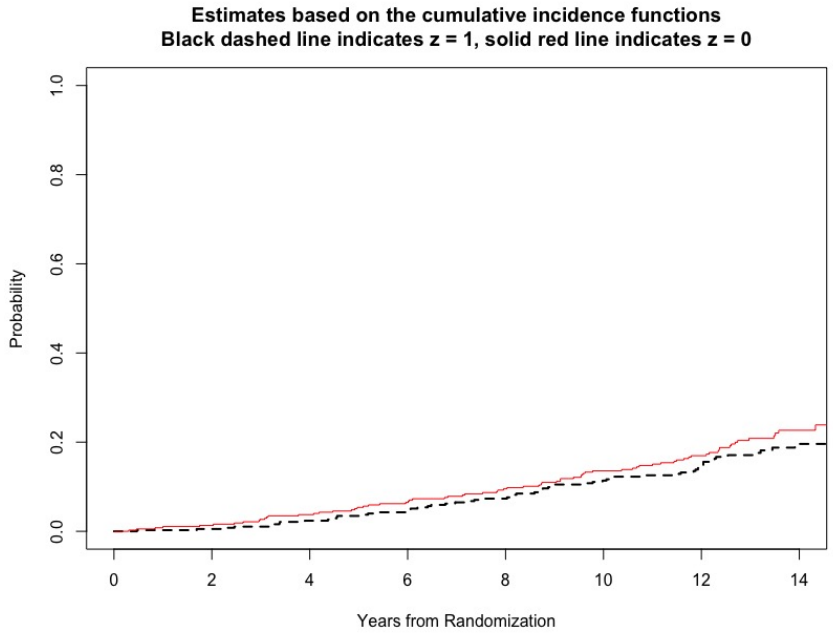


Figure 4.5: Cumulative incidence curves for the two treatment groups in the prostate cancer clinical trial.

4.6.1 Conventional Models

In the data, we consider the $z = 1$ group to be the treatment group for salvage radiation therapy with antiandrogen therapy, and the $z = 0$ represents the group treated without antiandrogen therapy. There is a significant treatment effect of the additional antiandrogen therapy on time to distant metastasis using a parametric hazard model with a Weibull baseline hazard ($HR = 0.622, p = 0.004$). The median survival time to S for the $z = 0$ arm and $z = 1$ arm is not reached. There is a marginally significant treatment effect on overall survival when considering the cause-specific hazard ($HR = 0.722, p = 0.049$). The median survival times for OS in the two arms are 15.7 and 16.3 years, respectively. Based on the Kaplan Meier curves and typical survival times, we chose $\tau_S = 5$ and τ_T = 8. We calculate the number of individuals who go through each transition and experience the events in our illness-death models. In total, 156 patients experienced distant

metastases, and 239 total deaths were observed between the two arms. These numbers are shown in Figure 4.6.

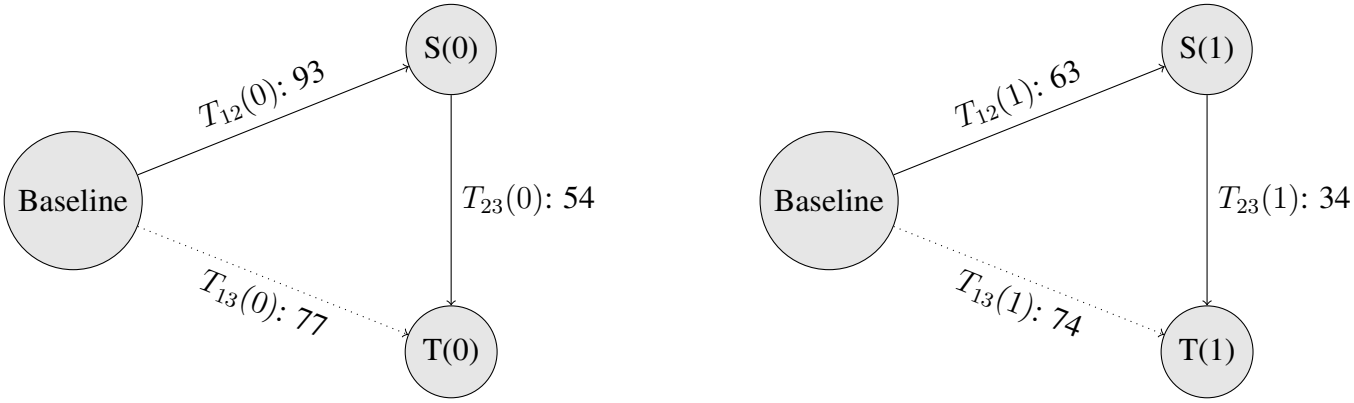


Figure 4.6: Counterfactual Illness-Death Models for baseline, illness (S), and death (T) with the number of individuals experiencing the events in each transition for the prostate cancer trial.

4.6.2 Surrogacy Evaluation

First we perform the analysis marginally. Here we show an estimated CEP curve based on several assumptions: the baseline hazard follows an exponential distribution, and we use model A using T_{12} as a time-varying covariate where we assume $\kappa_{12}^z = \kappa_{13}^z = \kappa_{23}^z = 1$. Table 4.5 shows the posterior mean and corresponding 95% credible interval for each parameter being estimated. We plot the posterior mean of ΔS_i and ΔT_i for each individual across iterations in a CEP plot. We also show the estimated slope and intercept lines on the CEP curve for each iteration of the MCMC chain to assess the variability of the estimates of these validation quantities.

Based on this example dataset and CEP curve without covariates in Figure 4.7, the 95% credible interval for the intercept term γ_0 is (-0.152, 0.080) with posterior mean -0.036. For the slope γ_1, the 95% credible interval is (0.017, 0.135) with posterior mean 0.076. Based on these estimates, we would conclude that the slope γ_1 is positive and the estimated intercept γ_0 is near zero since the

credible interval for γ_0 does include 0. These results would indicate that the surrogate seems valid, though the credible interval for γ_0 is somewhat wide. We also conducted a sensitivity analysis in a similar way that was described in the simulation studies. Instead of assuming $\omega_{13}^z = \omega_{23}^z$ and that $\omega_{12}^z \perp \omega_{13}^z$, we assumed that all six counterfactual frailties were correlated within an individual. These results gave reasonably similar conclusions, with an estimated γ_0 of -0.046 (-0.157, 0.073) and estimated γ_1 of 0.108 (0.045, 0.195).

Next we fit conditional surrogacy validation models. We include baseline PSA, age, and Gleason score as baseline covariates. It is likely that controlling for covariates will change the estimated frailties, as the frailty terms capture the unexplained heterogeneity in treatment effects. Using complete cases for covariates results in a sample size of 756 men, and we use each covariate in each transition. Based on this, we see that the estimated γ_0 is again near zero with a positive estimate of γ_1 in Figure 4.8. We find that the estimated quantities are 0.026 (-0.072, 0.124) for γ_0 and 0.050 (0.029, 0.072) for γ_1. Based on these analyses, we could also determine if the surrogate is valid for certain subgroups of people (Roberts et al. 2021). It is also possible that different covariates may be more important in different transition models. For example, we may expect age to be more important for the direct transition from baseline to death, while baseline PSA and Gleason score will likely be more important for time to distant metastases. Model selection could lower the number of parameters to estimate.

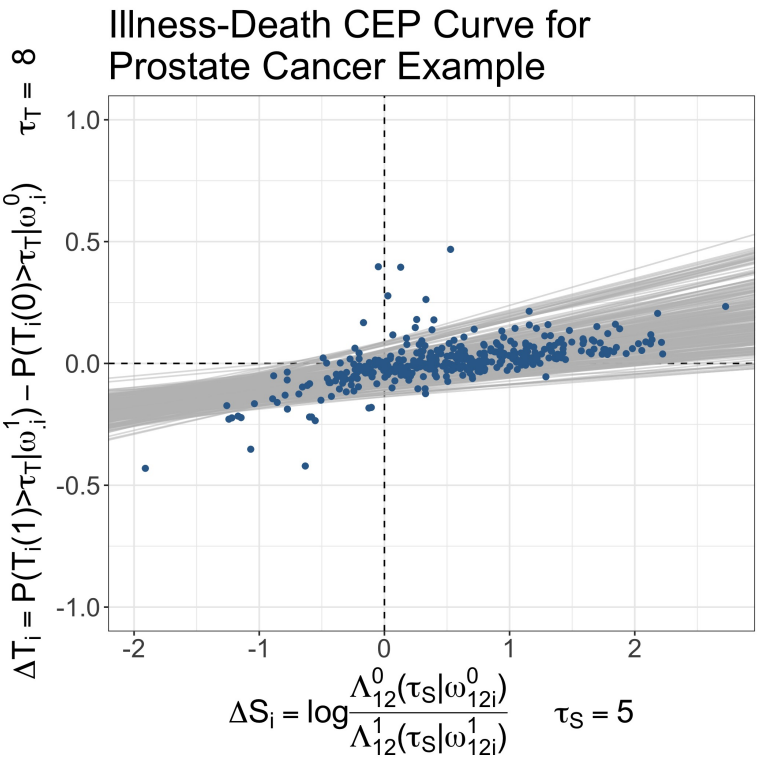


Figure 4.7: Causal effect predictiveness plot for the motivating prostate cancer trial dataset. Each point represents the posterior mean of ΔS_i and ΔT_i for an individual. The collection of linear best fit lines in gray represent the posterior slope γ_1 and intercept γ_0 evaluated at each iteration of the MCMC. No covariates are considered in this model.

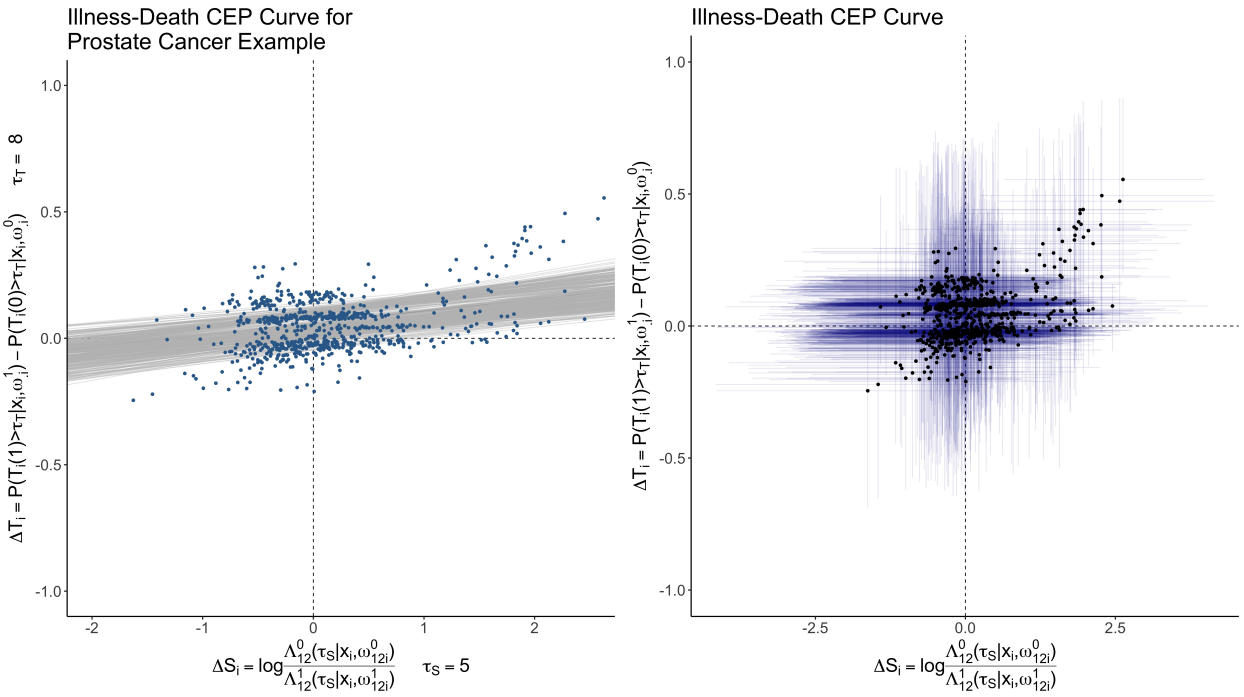


Figure 4.8: Causal effect predictiveness plot for the motivating prostate cancer trial dataset including the covariates age, PSA at baseline, and Gleason score as categorical covariates. Each point represents the posterior mean of ΔS_i and ΔT_i for an individual. On the left, the collection of linear best fit lines in gray represent the posterior slope γ_1 and intercept γ_0 evaluated at each iteration of the MCMC. On the right, each blue uncertainty interval represents the credible interval associated with each individual and their coordinate on the plot.

4.7 Discussion and Future Work

In this work, we have considered how to validate surrogate endpoints when trial outcomes are time-to-event using principal stratification and illness-death models. We have provided examples and an online app to explore CEP curves under different data settings. While the values of the CEP curve can be written in a closed, analytic form when the outcomes are Gaussian in previous chapters (Conlon et al., 2014a; Roberts et al., 2021), it is necessary to define and empirically assess what an ideal CEP curve looks like for time-to-event data. A novel distinction in this chapter is that in the Gaussian case, the CEP conditions on $S_i(1) - S_i(0) = s$, where the conditioning is on

a contrast between potentially observable values, $S_i(1)$ and $S_i(0)$. In this paper, we are looking at the contrast between Λ^z_{12i} and Λ^{1-z}_{12i}, which is a contrast between distributions.

While not the case in our considered scenarios, some extrapolation may be required to determine if the CEP curve goes through the origin of the plot depending on the size of the treatment effect on S. The subject-specific plotted points may not appear in all four quadrants of the plot. There is an interesting connection regarding individual specific ΔS_i and ΔT_i within the quadrants of the graph that has been considered when effects are plotted across trials in the meta-analytic setting (Elliott et al., 2015). In particular, certain subject-specific coordinates may suggest that the treatment has a beneficial effect on the surrogate endpoint but a detrimental effect on the true outcome for certain individuals. This may be informative when considering the possibility of the surrogate paradox (VanderWeele, 2013).

There are several areas for sensitivity analyses and exploration of identifiability for surrogacy validation (Ghosh, 2012). While the variance of the frailty should be identifiable by including sufficient covariates (Gao, 2012; Putter et al., 2015), it may still be difficult to accurately estimate frailty terms in a complex model. In our proposed models, we include a prior distribution for the variance of the frailty terms but do not assume the variance is known. Since allowing for too much flexibility in the models may result in non-identifiability of parameters, this can lead to identifiability problems when trying to estimate the coefficients associated with the frailties. We believe our assumptions that $\kappa_{jk} = 1$ or that $\omega^z_{13} = \omega^z_{23}$ about the frailty terms are justifiable for this data example. They also help with computation during estimation, but they are still potentially strong assumptions. Relaxing the assumption that the frailties going into the T state are equal (ie $\omega^z_{13} = \omega^z_{23}$) may impact identifiability since there will be less information available to estimate these terms. To the extent that frailties can be estimated for one event time per person, the data might inform these assumptions (e.g., the assumption is testable to the extent that frailties can be

estimated well). We might try to assess the identifiability of frailty terms in the proposed causal model by comparing the prior and posterior distributions for the frailty terms (Gao, 2012). Other convergence metrics can be used to assess the convergence of the parameters, and more complex algorithms or different distributional assumptions about the frailties may alleviate computational problems (Clayton, 1991; Wen et al., 2016 for example). For assessment of robustness, our models can be evaluated under model misspecification. To increase the flexibility of the method, we could also consider fitting a non-linear loess curve through the points on the CEP plot as opposed to a linear fit. We can compare our proposed methods to copula models (Taylor et al., 2015). These particular Gaussian copula models have potential of extending the closed-form correlation structure we have focused on in previous chapters while incorporating conditional independence assumptions on the appropriate correlation scale.

In the future, we can consider changing our model parameterization from our proposal to use a time-varying covariate in the transition model from S to T to the alternative Model B or a different structure. We may extend beyond the proposed illness-death model to a different or more complex multi-state model depending on the endpoints being evaluated. In different disease areas, consideration about individuals being cured may be appropriate (Conlon et al, 2014b). We have assumed here that time to S is known, but it may be subject to interval censoring. In some cases we may even have exact information about time to T based on death registries without knowing if S occurred (Beesley et al., 2019). Different models, definitions of the endpoint, and corresponding ΔS_i may change our determination whether the surrogate is valid, and the assumptions made about the models and frailties may be more appropriate for certain contexts.

It would be interesting to evaluate this illness death model when ΔS_i is based on a composite endpoint with T. For example, in the prostate cancer setting, distant metastases-free survival has been considered as a surrogate endpoint for overall survival. Other potential surrogates have been

considered such as biochemical recurrence or time to local metastases, and an alternative true clinical outcome could be prostate cancer-specific survival. In our setting, it is likely that individuals may only die from prostate cancer if they experience distant metastases, so there may be fewer individuals transitioning directly from baseline to cancer-specific death compared to baseline to death from other causes. It is interesting to consider how our mechanistic approach to disease progression explored by the illness-death models may compare to other techniques such as meta-analysis for surrogacy validation. For example, we may believe that our approach will assess surrogates in a way that is more generalizable across treatments than methods that rely on composite endpoints. This comparison and concept of transportability remains as future work (Pearl and Bareinboim, 2011).

There are several other directions for extending this work, particularly when considering the overlap of causal inference and survival analysis and delicate interpretation of hazard ratios with multiple time-to-event endpoints. Gran et al. (2015) explores other causal tools for multi-state models such as inverse probability weighting, G-computation, and manipulating hypothetical transition intensities. Other directions for future work are to formally compare the proposed models with the similar structures of the Prentice criteria, models using mediation strategies, or other causal methods.

4.8 Tables

	$\lambda_{12}^0 = \lambda_{12}^1$	$\lambda_{13}^0 = \lambda_{13}^1$	$\lambda_{23}^0 = \lambda_{23}^1$	Surrogacy
Scenario 1	T	T	T	Null Case
Scenario 2	F	T	T	Perfect
Scenario 3	F	T	F	Partial
Scenario 4	F	F	T	Partial
Scenario 5	F	F	F	Partial
Scenario 6	T	F	F	Not a surrogate
Scenario 7	T	T	F	Not a surrogate
Scenario 8	T	F	T	Not a surrogate

Table 4.1: Eight possible scenarios of which pathways in the illness-death models exhibit treatment effects based on the causal hazards. T denotes true and F denotes false. The right-hand column represents an intuitive notion of whether S is a good surrogate for T.

	γ_0	γ_1
Scenario 1: True Value*	-0.062	0.090
Estimates	-0.061	0.089
SE	0.031	0.011
SD	0.029	0.008
Scenario 2: True Value*	-0.043	0.093
Estimates	-0.036	0.091
SE	0.023	0.007
SD	0.024	0.006
Scenario 3: True Value*	-0.020	0.081
Estimates	-0.017	0.080
SE	0.033	0.011
SD	0.032	0.008
Scenario 4: True Value*	-0.029	0.103
Estimates	-0.025	0.105
SE	0.033	0.011
SD	0.032	0.008
Scenario 5: True Value*	0.037	0.091
Estimates	0.044	0.091
SE	0.034	0.011
SD	0.034	0.008
Scenario 6: True Value*	0.035	0.086
Estimates	0.049	0.085
SE	0.032	0.011
SD	0.032	0.008
Scenario 7: True Value*	0.007	0.078
Estimates	0.010	0.077
SE	0.032	0.011
SD	0.030	0.008
Scenario 8: True Value*	-0.035	0.098
Estimates	-0.025	0.099
SE	0.031	0.012
SD	0.031	0.008

Table 4.2: Simulation results from illness-death models and estimated validation quantities. This table shows the posterior mean, average estimated standard error (SE), and the standard deviation (SD) of the posterior means across simulation replications.
In these calculations, the κ_{jk} parameters are fixed.
*Based on empirical calculations from a larger sample size over many replications

	γ_{12}^0	γ_{13}^0	γ_{23}^0	γ_{12}^1	γ_{13}^1	γ_{23}^1	θ_{23}^0	θ_{23}^1
Scenario 1: True Value*	1	0.5	1	1	0.5	1	0	0
Estimates	0.93	0.47	0.97	0.93	0.48	0.98	0.004	0.008
SE	0.07	0.04	0.08	0.07	0.04	0.08	0.059	0.058
SD	0.07	0.04	0.08	0.07	0.04	0.10	0.086	0.082
Scenario 2: True Value*	1	0.5	1	0.61	0.5	1	0	0
Estimates	0.93	0.47	1.01	0.57	0.48	1.03	-0.066	-0.057
SE	0.07	0.04	0.09	0.05	0.04	0.10	0.059	0.056
SD	0.07	0.04	0.08	0.05	0.03	0.12	0.088	0.077
Scenario 3: True Value*	1	0.5	1	0.61	0.5	0.61	0	0
Estimates	0.93	0.47	1.03	0.57	0.48	0.64	-0.083	-0.075
SE	0.07	0.04	0.09	0.05	0.04	0.06	0.057	0.055
SD	0.07	0.04	0.08	0.05	0.03	0.07	0.088	0.073
Scenario 4: True Value*	1	0.5	1	0.61	0.31	1	0	0
Estimates	0.93	0.47	0.97	0.57	0.32	0.97	-0.002	0.011
SE	0.07	0.04	0.08	0.04	0.00	0.09	0.059	0.054
SD	0.07	0.04	0.08	0.04	0.04	0.12	0.085	0.076
Scenario 5: True Value*	1	0.5	1	0.61	0.31	0.61	0	0
Estimates	0.93	0.47	0.99	0.57	0.32	0.61	-0.023	-0.008
SE	0.07	0.04	0.08	0.04	0.04	0.07	0.084	0.070
SD	0.07	0.04	0.07	0.07	0.05	0.10	0.084	0.080
Scenario 6: True Value*	1	0.5	1	1	0.5	0.61	0	0
Estimates	0.93	0.47	0.94	0.93	0.36	0.58	0.040	0.049
SE	0.07	0.04	0.08	0.07	0.01	0.05	0.058	0.055
SD	0.07	0.04	0.07	0.06	0.05	0.06	0.081	0.074
Scenario 7: True Value*	1	0.5	1	1	0.5	0.61	0	0
Estimates	0.93	0.47	0.98	0.93	0.47	0.61	-0.021	-0.017
SE	0.07	0.04	0.08	0.07	0.04	0.05	0.059	0.058
SD	0.07	0.04	0.08	0.07	0.04	0.06	0.084	0.075
Scenario 8: True Value*	1	0.5	1	1	0.31	1	0	0
Estimates	0.94	0.47	0.93	0.93	0.36	0.93	0.065	0.072
SE	0.07	0.04	0.08	0.07	0.01	0.08	0.056	0.053
SD	0.07	0.04	0.07	0.07	0.05	0.10	0.084	0.080

Table 4.3: Simulation results from illness-death models of Weibull distribution parameters γ and coefficients θ_{23}.
*The true value of γ_{12}^1, γ_{13}^1, and γ_{23}^1 depends on the scenario.
In these calculations, the κ_{jk} parameters are fixed, $\omega_{13}^z = \omega_{23}^z$, we estimate θ_{23}^z.

	γ_0	γ_1	γ_{12}^0	γ_{13}^0	γ_{23}^0	γ_{12}^1	γ_{13}^1	γ_{23}^1
Frailties Independent: True Value*	0.000	0.039	1	0.5	1	0.61	0.5	1
Estimates	0.014	0.030	0.98	0.49	0.94	0.60	0.49	0.96
SE	0.076	0.017	0.07	0.05	0.26	0.05	0.04	0.33
SD	0.025	0.008	0.07	0.04	0.10	0.05	0.04	0.11
Frailties Dependent: True Value*	-0.009	0.083	1	0.5	1	0.61	0.5	1
Estimates	-0.022	0.085	0.98	0.49	0.89	0.60	0.49	0.90
SE	0.091	0.019	0.07	0.05	0.29	0.05	0.04	0.36
SD	0.025	0.008	0.07	0.04	0.09	0.05	0.04	0.10

Table 4.4: Simulation results from illness-death models of Weibull distribution scale parameters γ_{jk} and regression coefficients θ_{23}. We compare the results when making different assumptions about the frailties: either $\omega_{12}^z \perp \omega_{13}^z \perp \omega_{23}^z$ or they are all correlated. In either case, they are unequal.

* Scenario 2 shown here. In these calculations, the κ_{jk} and α_{jk} parameters are fixed, and $\tau_S = 1, \tau_T = 2$.

Parameter	γ_0	γ_1	γ_{12}^0	γ_{13}^0	γ_{23}^0	γ_{12}^1	γ_{13}^1	γ_{23}^1	θ_{23}^0	θ_{23}^1
Marginal (no covariates)										
Posterior Mean	-0.036	0.076	0.018	0.018	0.172	0.013	0.015	0.266	0.097	0.035
SE	0.059	0.030	0.002	0.002	0.180	0.002	0.002	0.371	0.248	0.243
Parameter	γ_0	γ_1	γ_{12}^0	γ_{13}^0	γ_{23}^0	γ_{12}^1	γ_{13}^1	γ_{23}^1	θ_{23}^0	θ_{23}^1
Including covariates										
Posterior Mean	0.026	0.050	0.008	0.024	0.348	0.012	0.017	0.385	-0.033	-0.040
SE	0.050	0.011	0.004	0.003	0.047	0.002	0.004	0.070	0.017	0.020
Parameter	$\beta_{gleason12}^0$	β_{psa12}^0	β_{age12}^0	$\beta_{gleason12}^1$	β_{psa12}^1	β_{age12}^1				
Including covariates										
Posterior Mean	0.046	0.623	-0.045	0.314	0.633	-0.001				
SE	0.072	0.021	0.028	0.024	0.030	0.014				
Parameter	$\beta_{gleason13}^0$	β_{psa13}^0	β_{age13}^0	$\beta_{gleason13}^1$	β_{psa13}^1	β_{age13}^1				
Including covariates										
Posterior Mean	0.121	0.483	0.214	0.418	0.634	0.381				
SE	0.014	0.025	0.018	0.036	0.012	0.097				
Parameter	$\beta_{gleason23}^0$	β_{psa23}^0	β_{age23}^0	$\beta_{gleason23}^1$	β_{psa23}^1	β_{age23}^1				
Including covariates										
Posterior Mean	0.222	0.275	0.329	0.233	0.082	0.518				
SE	0.039	0.017	0.019	0.008	0.023	0.016				

Table 4.5: Parameter estimates for the prostate cancer data example. The posterior mean and estimated standard error are shown for each parameter. All α_{jk} and κ_{jk} are set to 1.

CHAPTER V

Conclusion

This dissertation develops methods that can be applied to randomized clinical trials using the causal inference framework. Intermediate endpoints can serve as surrogates for a true clinical outcome and improve the efficiency of the trial. The ability for a trial to run in shorter time could speed up drug approval and help patients. As we have discussed in the previous chapters, valid assessment as to whether a proposed surrogate endpoint is appropriate to use in a future trial is challenging. It is very costly and potentially dangerous to wrongly claim a treatment benefit based on a biologically-lacking or otherwise inadequate endpoint (Vanderweele, 2013). The development of rigorous trial protocols and methodology is crucial for the approval of beneficial treatments and success of drug-development. This dissertation extends causal association approaches to validate a candidate surrogate outcome using potential outcomes. We give attention to a range of outcome types that are applicable in clinical trials for essentially any disease type. In this chapter, we summarize the methods proposed in chapters II, III, and IV, and explore directions to continue this work in the future.

The proposed surrogate validation methods are based on the principal stratification framework (Frangakis and Rubin, 2002), where we jointly model the potential outcomes of the surrogate S and true clinical endpoint T under a binary treatment. In chapters II and III, our approach is motivated by an ongoing study of a muscular dystrophy gene therapy. The candidate surrogate

is an expression of micro-dystrophin that is measured a few weeks after the therapy is initiated. One characteristic of the disease is a lack of protein due to micro-dystrophin, and subjects with the disease and without gene therapy will have essentially zero gene expression. The gene therapy is aiming to activate this gene. Therefore this setting of interest allows us to assume the surrogate under the placebo, $S(0)$, is zero-valued, and ideally the treatment would increase the value of $S(1)$. This is a simplification we can reasonably make for this trial and is known as the constant biomarker assumption (Gilbert and Hudgens, 2008). Based on this setting, we develop methods to incorporate conditional independence and other modeling assumptions to explore their impact on the assessment of surrogacy. We compare the estimation properties of a Bayesian imputation method using Markov Chain Monte Carlo to strategies using the observed data only, and we explore the impact of different prior distributions on non-identified parameters.

Chapter III also considers this motivating clinical trial for muscular dystrophy, where the outcomes are measured longitudinally. We develop a mixed model approach that can potentially gain estimation efficiency by modeling the repeated measures of T via random intercepts or random slopes. Further, it may be possible to measure additional T and S outcomes in a delayed-treatment start trial design. In this situation, subjects who are first administered the placebo may be given the gene therapy mid-trial. This design would aid with identifiability of model parameters, and we extend our models and metrics for validation in such a trial. This chapter also proposes novel conditional independence assumptions of counterfactual random effects. Lastly, we also consider how to define the quantities for validation such that they may depend on time. It is plausible that S may only be valid within a certain time proximity that T is measured.

In Chapter IV, we extend these ideas to the surrogate validation framework with time-to-event data for both the surrogate marker and the final outcome of interest. This setting becomes more complex as S and T are not guaranteed to be observed during the study period. We develop

a method that incorporates the censoring and semi-competing risk structure that is likely to be encountered with multiple survival endpoints. We maintain a valid causal interpretation by viewing this through the lens of a causal illness-death (multi-state) model. In this chapter, we extend our proposal of subject-specific random effects using frailty terms. We investigate how to define the parameters measuring the association between outcomes and relevant principal strata and quantify which settings would result in a valid surrogate. Finally, we demonstrate our method on data from a prostate cancer clinical trial.

A large component of this dissertation examines how to address the non-identified parameters in the proposed causal models. Both chapters II and III assess how imputation algorithms with the complete data likelihood compare to methods that use only the observed data. This is important as we expect the constrained prior and posterior distributions to be equal for non-identified correlation parameters once we account for constraints on the covariance matrix. The imputation scheme can be computationally burdensome and may encounter converge issues. Estimation of random effects based on these non-identified parameters relies upon an additional layer of distributional assumptions. Bayesian methods allow for the incorporation of prior distributions on these correlation parameters to help with identifiability. Sensitivity analyses are also a popular technique that involve fixing the parameters and testing to what extent the results change over different fixed values. Existing literature elaborates more on these topics in terms of transparent parameterizations, bounds, and sensitivity analysis (Gustafson, 2010; Richardson et al., 2010). These ideas are relevant beyond the surrogate endpoint setting for other causal methods.

There are clear directions to make this work more general when we have non-Gaussian outcomes. Generalized linear mixed models may be used to continue ideas from Chapters II and III to make the methods applicable in more settings. Conceptually, extensions with non-linear link functions are straightforward, though the computational complexity of estimating non-Gaussian

random effects needs to be investigated. Another case we could extend this work would be where $S(0)$ is not necessarily equal to 0. The methods that are presented in Chapter II are all easily extendable to this case. With covariates, the quantities for γ_0 and γ_1 can be derived from the conditional distribution of $T(1)-T(0)|S(1)-S(0), X$ which is based upon a larger multivariate normal distribution.

$$\begin{bmatrix} S_i(0) \\ S_i(1) \\ T_i(0) \\ T_i(1) \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mu_{S_0} + \psi_1 X_i \\ \mu_{S_1} + (\psi_1 + \psi_2) X_i \\ \mu_{T_0} + \omega_1 X_i \\ \mu_{T_1} + (\omega_1 + \omega_2) X_i \end{bmatrix}, \begin{bmatrix} \sigma^2_{S_0} & \rho_S \sigma_{S_0} \sigma_{S_1} & \rho_{00} \sigma_{S_0} \sigma_{T_0} & \rho_{01} \sigma_{S_0} \sigma_{T_1} \\ & \sigma^2_{S_1} & \rho_{10} \sigma_{S_1} \sigma_{T_0} & \rho_{11} \sigma_{S_1} \sigma_{T_1} \\ & & \sigma^2_{T_0} & \rho_t \sigma_{T_1} \sigma_{T_0} \\ & & & \sigma^2_{T_1} \end{bmatrix} \right)$$

$$E(T(1) - T(0)|S(1) - S(0) = s, X = x) =$$

$$(\mu_{T1} - \mu_{T0}) - (\frac{\rho_{11}\sigma_{S1}\sigma_{T1} - \rho_{10}\sigma_{S1}\sigma_{T0} - \rho_{01}\sigma_{S0}\sigma_{T1} + \rho_{00}\sigma_{S0}\sigma_{T0}}{\sigma^2_{S0} + \sigma^2_{S1} - 2\rho_s\sigma_{S0}\sigma_{S1}})(\mu_{S1} - \mu_{S0})$$

$$+(\frac{\rho_{11}\sigma_{S1}\sigma_{T1} - \rho_{10}\sigma_{S1}\sigma_{T0} - \rho_{01}\sigma_{S0}\sigma_{T1} + \rho_{00}\sigma_{S0}\sigma_{T0}}{\sigma^2_{S0} + \sigma^2_{S1} - 2\rho_s\sigma_{S0}\sigma_{S1}})s+$$

$$(\omega_2 - (\frac{\rho_{11}\sigma_{S1}\sigma_{T1} - \rho_{10}\sigma_{S1}\sigma_{T0} - \rho_{01}\sigma_{S0}\sigma_{T1} + \rho_{00}\sigma_{S0}\sigma_{T0}}{\sigma^2_{S0} + \sigma^2_{S1} - 2\rho_s\sigma_{S0}\sigma_{S1}})\psi_2)x$$

$$= \gamma_0 + \gamma_1 s + (\omega_2 - (\frac{\rho_{11}\sigma_{S1}\sigma_{T1} - \rho_{10}\sigma_{S1}\sigma_{T0} - \rho_{01}\sigma_{S0}\sigma_{T1} + \rho_{00}\sigma_{S0}\sigma_{T0}}{\sigma^2_{S0} + \sigma^2_{S1} - 2\rho_s\sigma_{S0}\sigma_{S1}})\psi_2)x$$

In this more general setting, there are more non-identified correlation parameters. We can explore different conditional independence assumptions given X. From the distribution of $T(1) - T(0)|S(1) - S(0)$, possible assumptions and corresponding constraints can be written in a relative order from most to least restrictive

1. $T(1) \perp T(0)|X, S(0), S(1) \Rightarrow \rho_t = \frac{\rho_{11}\rho_{10}+\rho_{01}\rho_{00}-\rho_s(\rho_{01}\rho_{10}+\rho_{11}\rho_{00})}{(1-\rho_s^2)}$

2. $S(1) \perp T(0)|X, S(0), T(1)$

3. $S(0) \perp T(1)|X, S(1), T(0)$

4. $S(1) \perp T(0)|X, S(0)$

5. $S(0) \perp T(1)|X, S(1)$

$$2 \Rightarrow \frac{\rho_S\rho_{00} - \rho_{11}\rho_{01}\rho_{00} + \rho_t\rho_{11} - \rho_S\rho_t\rho_{01}}{\rho_{10}(1-\rho_{01}^2)} = \frac{\sigma_{S1}^2\sigma_{T0}^2}{\sigma_{S0}^2\sigma_{T1}^2}$$

$$3 \Rightarrow \frac{\rho_S\rho_{11} - \rho_{00}\rho_{10}\rho_{11} + \rho_t\rho_{00} - \rho_S\rho_t\rho_{10}}{\rho_{01}(1-\rho_{10}^2)} = \frac{\sigma_{S0}^2\sigma_{T1}^2}{\sigma_{S1}^2\sigma_{T0}^2}$$

$$4 + 5 \Rightarrow \frac{\rho_{01}}{\rho_{11}} = \frac{\rho_{10}}{\rho_{00}} = \rho_s$$

We can also reparameterize the model in the following way

$$S(0), S(1)|X \sim N(\begin{bmatrix} \kappa_0 + \nu_1 X \\ \kappa_1 + \nu_2 X \end{bmatrix}, \begin{bmatrix} \tau_1^2 & \pi \\ \pi & \tau_2^2 \end{bmatrix})$$

$$T(0), T(1)|S(0), S(1), X \sim N(\begin{bmatrix} \theta_0 + \delta_1 X + \eta_1 S0 \\ \theta_1 + \delta_2 X + \eta_2 S1 \end{bmatrix}, \begin{bmatrix} \tau_3^2 & \lambda \\ \lambda & \tau_4^2 \end{bmatrix})$$

for the constraint corresponding to

$$S(1) \perp T(0)|X, S(0) \text{ and } S(0) \perp T(1)|X, S(1)$$

Due to the conditional independence assumption, two coefficient values are 0. Now there are only two non-identified covariances, λ and π. Accommodating these parameters to ensure positive definite covariance matrices may be more simple than considering a four-dimensional Gaussian distribution.

Other possibilities to extend the models include when S is repeatedly measured. In this case, we

would extend the models proposed in chapter III to include random effects for $S(1)$ and potentially $S(0)$. In the case of random intercept models, this would lead to a four-dimensional multivariate normal distribution similar to what was described above. With the longitudinal models, there may be value in further considering the individual-level trajectories based on the random effects. For example, we could interpret the area between the counterfactual, individual-specific outcome curves over time. Alternatively, it may also be of interest to see how generalized estimating equations may be implemented and interpreted at a population level instead of the mixed models.

Outside of surrogate validation, some existing methods to estimate causal effects use Bayesian non-parametrics or Dirichlet process priors (Xu et. al, 2020). We could incorporate these semi- or non-parametric methods in our framework for more flexibility. Alternatively, copula models show promise for continuing the Gaussian-based correlation structure we have relied on in chapters II and III. Conlon et al. (2017a) proposes a method for non-Gaussian endpoints that involves imputation of censored survival times and transforming the ordinal and survival times to a joint normal distribution. They estimate the correlation parameters among the latent, Gaussian variables and impute missing counterfactual outcome values before transforming back to the original survival and ordinal variable scale to calculate the validation metrics.

It is not readily apparent at which steps to include covariates in the model and on which scale to invoke the conditional independence assumptions we have considered thus far. It seems reasonable that the marginal distribution of the survival times could depend on covariates. Many authors recommend using covariates in regression models on the original scale to preserve a marginal interpretation of the coefficients (Song, 2009; Masarotto and Varin, 2012). Alternatively, the conditional independence conditions we have been considering on the correlation matrix would be applicable on the joint Gaussian scale. Pitt (2006) describes a Bayesian method for estimating a copula regression model with conditional independence assumptions invoked on the correlation

matrix rather than among the data that are actually observed (Bhadra et al., 2018). However, the metrics used for validating the surrogate, such as the CEP surface, are assessed on the observed (non-transformed) scale. Clarifying which scale to make these assumptions on is an area for further exploration.

A major topic involved in Chapter IV is how to obtain causal interpretations for time-to-event outcomes. One reason we propose causal illness-death models is that it is not clear what appropriate causal estimand can incorporate the risk sets of individuals who may experience S or T when we want to condition on strata of individuals. Briefly, we explore some of the related issues suggested about the commonly used hazard ratio and the difficulty in assigning it a causal interpretation (Hernán, 2010). Since the population being assessed for a treatment effect becomes smaller as events occur and the balance of the study population is lost, the hazard ratio may change over time. That is, groups of individuals who survive to some $t > 0$ with and without treatment, $T(1) \geq t, T(0) \geq t$ will not be comparable if treatment affects the outcome since susceptibility is only randomized at baseline. A selection bias appears due to this frailty effect, which has been formalized by demonstrating that conditioning on survival to time $t > 0$ leads to a collider bias (Aalen, Cook, and Røysland, 2015). Groups surviving past t with or without treatment will be comparable if $T(1) \perp T(0)|Z$ for any confounders Z. The related problem of unmeasured covariates in a Cox proportional hazards model has been long acknowledged (Henderson and Oman, 1999; Omori and Johnson, 1993). It is not clear if a simple frailty model will sufficiently address this phenomena and to obtain a causal interpretation without considering potential outcomes. A principal strata approach has been proposed to address this concern, including the conditional hazard ratio $\frac{\lim_{h\to 0} P(t\leq T(1)<t+h|T(0)\geq t,T(1)\geq t)}{\lim_{h\to 0} P(t\leq T(0)<t+h|T(0)\geq t,T(1)\geq t)}$ (Martinussen et al., 2020). This is based upon the principal stratum of individuals who would have survived up to time t regardless of treatment. It would be interesting to extend or compare our work in Chapter IV with principal strata under a

multi-state model with frailty terms to these ideas. Methods could be borrowed across the topics to obtain interpretable and causal estimands with for time-to-event data with or without intermediate endpoints.

Other clinical trial designs using intermediate outcomes could be considered and integrated in future work. We could consider using intermediate markers for the related purpose of defining a stopping rule for futility in a trial (see Parmar et al., 2008 and others for examples where this futility marker is different from the true outcome). Building upon our notation in this dissertation, for treatment $Z = z$, let $S(z)$ denote the intermediate outcomes (though not necessarily validated surrogates), $T(z)$ denote the clinical outcomes of interest, and θ define the observable treatment effect on S. The trial may be stopped early if $\hat{\theta} < k$ for some predetermined threshold k. Most existing, rigorous rules for choosing the stopping rule threshold are based on error spending functions (Maurer and Bretz, 2013) as opposed to causal concepts and potential outcomes. In particular, many are based on error rates for the assumed distribution of the intermediate outcome (Sydes et al., 2009) which does not directly incorporate the relationship between S and T. The value of such a futility marker would depend largely on the correlation between the marker and the true outcome, which suggests some connection to the surrogate validation framework that could be investigated with a causal inference approach.

For an individual i, θ_i is defined as some contrast between $S_i(1)$ and $S_i(0)$ where we would continue the trial when both $S_i(1) > S_i(0)$ and $T_i(1) > T_i(0)$ and stop early when both $S_i(1) < S_i(0)$ and $T_i(1) < T_i(0)$. Wrongly stopping a beneficial treatment early is similar to a type II error: $S_i(1) < S_i(0)$ but $T_i(1) > T_i(0)$, and the opposite type I-like error occurs when the trial wrongly continues: $S_i(1) > S_i(0)$ but $T_i(1) < T_i(0)$. For surrogacy, the two conditions of average causal necessity and average causal sufficiency are required, ensuring that there is no average effect of the treatment on T where there is no average effect on S, and similarly that an average

treatment effect on T exists where is there an average effect on S, respectively. Even if a candidate surrogate does not pass stringent surrogate criteria, it can still serve as a helpful auxiliary variable to improve inference on T (Li and Taylor, 2010), and thus S may still be informative for futility. For stopping rules, we may unequally more concerned about making a type II error, which is a concept not currently addressed in the principal surrogacy quantities. One possibility is that causal criteria could be developed on the individual level to define a rule with these desired properties and operating characteristics.

It seems natural to leverage information from completed trials to determine first, if an intermediate endpoint could serve as the stopping rule marker, and secondly the value that the corresponding cutoff k should take. Several methods for surrogate and biomarker validation are based on correlative measures that can be identified from multiple trials. The meta-analytic framework for surrogacy with mixed-effects models is based on the joint distribution of treatment effects for trial $i = 1, ..., n$ and individual $j = 1, ..., m$ (Burzykowski and Buyse, 2006)

$$S_{ij} = (\mu_S + m_{Si}) + (\alpha + a_i)Z_{ij} + \epsilon_{Sij} \qquad T_{ij} = (\mu_T + m_{Ti}) + (\beta + b_i)Z_{ij} + \epsilon_{Tij}$$

$$\begin{pmatrix} m_{Si} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} d_{ss} & d_{st} & d_{sa} & d_{sb} \\ & d_{tt} & d_{ta} & d_{tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix} \right)$$

$$\begin{pmatrix} \epsilon_{S_{ij}} \\ \epsilon_{T_{ij}} \end{pmatrix} \sim BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix} \right)$$

The surrogate validation quantity R^2_{trial} is obtained by regressing b_i on m_{Si} and a_i, and an R^2_{trial} value close to one represents a highly valid surrogate. Further, for a valid surrogate the treatment

effects on S and T must be correlated across trials with both null and nonzero treatment effects. In futility analysis, we should also consider how the treatment effects are associated at the trial level.

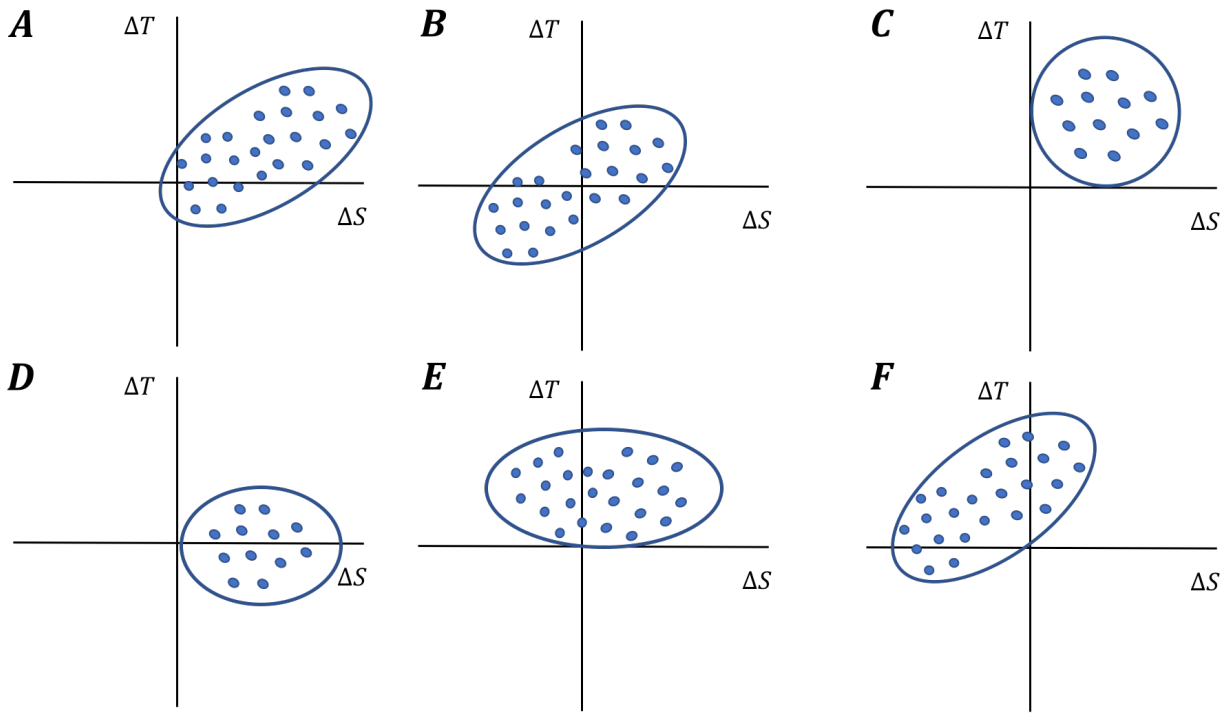


Figure 5.1: Potential associations between the treatment effects, Δ_S and Δ_T, for different treatment effects. While in most scenarios the two outcomes are correlated, that fact alone is not sufficient to determine whether S is a good marker and if a trial should stop for futility.

Consider the effects in Figure 5.1 where the hazard ratio for T is on the y-axis and the hazard ratio for S is on the x-axis. Elliott et al. (2015) explored regions of treatment effects within meta-analysis to identify the minimum, observed treatment effect for a surrogate S that will reduce the probability that the effect on T is harmful. A related goal for futility is to determine the minimum observed beneficial treatment effect for S corresponding to a high probability that the true treatment effect for T is null so that these trials can be stopped early. Based on the mixed model parameters above, Burzykowski and Buyse define the upper prediction limit function and corresponding surrogate threshold effect (STE). This estimates the minimum value of the treatment effect on S for which the predicted effect on T will be significantly different from zero. For

purposes of futility, we want to define a different threshold k such that if the treatment effect on S does not cross k, we are confident we will not observe a treatment effect on T. While related, this is distinct from the STE. Further, while these quantities are based on completed trials, Li and Taylor note that the data available mid-trial is S on a fraction of the n subjects and potentially also T on some subset of these. This is the information that would be available at an evaluation time point for futility and should be formally modelled in future methods for futility.

Another area of possible relation with our work is the use of intermediate endpoints within sequential multiple assignment randomized trials (SMART) (Murphy, 2005). In this design, also known as a dynamic treatment regime, intermediate outcomes are used to make adaptive treatment decisions. Examples of these decisions include dose or therapy modality at intermediate stages of the trial based on patient outcomes and tailoring functions. The sequence of decision rules is made on an individual patient basis with the goal of identifying the optimal adaptive treatment strategy. Potential outcomes could be used to reframe SMART trial designs under the causal inference lens, or we could develop methods and criteria to determine if the variable being used to make intermediate treatment decisions is deemed valid.

A final direction stemming from our discussion of integrating multiple trials and work from Chapter III with smaller sample sizes is the use of external data within surrogacy validation metrics. The use of historical data during trials is of increasing interest, and there is excitement about using electronic health records (EHR) or real world evidence and observational data to improve the generalizability and efficacy of trials. The first step toward incorporating arbitrary observational data for surrogacy is to create models that allow for external experimental or natural history data. We first consider how to merge these data sets with a small sample size randomized clinical trial to estimate a treatment effect and increase efficiency.

We consider the trial from Chapter III for a rare type of muscular dystrophy where the sam-

ple size may be too small to obtain a precise estimated treatment effect. We allow for repeated measures of the outcome which is the North Star Ambulatory Assessment (NSAA) functional measure including a cross-over portion and pre-randomization measurements of NSAA. With careful choice of the time scale (either age or time from randomization), we propose a combined model that is applicable to integrating three collected data sets (randomized data, natural history, and experimental data). From historical data, we define a model for the natural history of NSAA in untreated children and append this to the model for the effect of treatment on the longitudinal patterns of NSAA. For notation, let T_i be the outcome of NSAA measurement of subject i, B_i is age, A_i is age of subject i at the time of the gene therapy, and τ denotes time after randomization, so $A_i = B_i + \tau$. Let $H(\tau, A_i)$ be the treatment effect for subject i where $H(\tau, A_i) = E[T_i(\tau)|A_i, z_i = 1] - E[T_i(\tau)|A_i, z_i = 0]$. This represents the difference in the NSAA between the treated and untreated at time τ after treatment. To obtain an estimate of this quantity, we can model the trajectories for each data source with fixed effects X and random effects Z. Each subject has two pairs of random effects, an intercept and slope without treatment ($b_i^{(0)}$), and an intercept and slope with treatment ($b_i^{(1)}$)

$$\begin{pmatrix} b_i^{(0)} \\ b_i^{(1)} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} b_{0i}^{(0)} \\ b_{1i}^{(0)} \end{pmatrix} \\ \begin{pmatrix} b_{0i}^{(1)} \\ b_{1i}^{(1)} \end{pmatrix} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Psi} \right)$$

The model structure and proposed trajectory for each individual in the randomized trial is the same as we have proposed in Chapter III . At time of intervention, there is an immediate change in the fixed effect and a change in the slope for age.

Let θ_j for $j = 1, 2, 3$ be the population parameters for each data set, where $\theta_j \in \theta$, the set of

parameters. The likelihood from the randomized trial is $L_{RCT}(Y_R, \theta_1)$, from natural history study is $L_{HIST}(Y_H, \theta_2)$, and from the experimental study is $L_{EXP}(Y_E, \theta_3)$. Let $D = (Y_{RCT}, Y_H, Y_E)$ for all of the data so that $L(D, \theta) = L_{RCT} \times L_{HIST} \times L_{EXP}$. Several options are available for estimation, including maximum likelihood estimation or Bayesian estimation with potentially mildly informative priors. We may want to down-weight the external data, meaning $L_{RCT} \times (L_{HIST} \times L_{EXP})^{\alpha}$ for some parameter α. There are several methods questions to pursue in this setting. For example, the two arms of the randomized trial are likely to show differences due to the small sample size. It is possible that the subjects in the two additional data sets have different covariate distributions if they are generally more sick or healthy than the randomized patients. The natural history data may be less relevant if it is not current. One advantage of the proposed mixed model is that the timing and frequency of the T measurements can be different in the three data sets, and each data set may have its own random effects to account for these differences. Ultimately, this framework could also be extended to include the intermediate outcome being measured in at least one of the data sets to assess surrogacy.

APPENDIX A

Imputation Algorithm Details

For drawing the model parameters and potential outcomes, we consider the full data likelihood as follows, first considering the case with no covariates in general:

$$(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\begin{pmatrix} S(1)-\mu_{S1} \\ T(0)-\mu_{T0} \\ T(1)-\mu_{T1} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} S(1)-\mu_{S1} \\ T(0)-\mu_{T0} \\ T(1)-\mu_{T1} \end{pmatrix}\right)$$

As the conjugate, posterior distributions can be written in closed-form for most identified parameters, we use Markov Chain Monte Carlo (MCMC) methods. Let subscript l denote the l^{th} iteration of the Gibbs sampler, μ generally denote the set of mean parameters in the model and QRQ denote the set of variance and correlation parameters in the model.

$$\begin{array}{l|c|l|l} T_i(0) & S_i(1), T_i(1), & S_i(1) & T_i(0), \mu^{l-1}, \\ & \mu^{l-1}, R^{l-1}, Q^{l-1} & T_i(1) & R^{l-1}, Q^{l-1} \end{array}$$

During each iteration of the MCMC, we impute the missing potential outcome under treatment $z = 0$ for those who we observed an outcome under treatment $z = 1$ and vice versa. After the outcomes (denoted in general as Y) are imputed, we draw the mean, variance, and correlation parameters respectively:

For coefficients, $\mu|\cdot \sim$ Matrix Normal $\big((X^TX + \Lambda_0)(X^TY), X^TX + \Lambda_0, \Sigma\big)$ for prior matrix Λ_0

$\sigma_Y|\cdot \propto \sigma_Y^{-n}\exp(-\frac{1}{2}\Sigma_{i=1}^n(Y_i-\mu)(QRQ)^{-1}(Y_i-\mu)^T)$

For $j = T, 10, 11,\ \ \rho_j|\cdot \propto |R|^{-n/2} \exp(-\frac{1}{2}\Sigma_{i=1}^{n}(Y_i - \mu)(QRQ)^{-1}(Y_i - \mu)^T)$ for Uniform prior and within bounds determined by positive definiteness so the determinant is positive: $1-\rho_T^2+\rho_{10}^2 > \rho_{11}^2$. The condition that $\rho_{10} = \rho_T \times \rho_{11}$ is found by either setting a term in the precision matrix to 0 or solving for the covariance of $S(1), T(0)|T(1)$ and setting this equal to 0. Since the outcomes are multivariate normal, the conditional covariance is equal to

$$\begin{pmatrix} \sigma_{S1}^2 - \sigma_{S1}^2\rho_{11} & \rho_{10}\sigma_{T0}\sigma_{S1} - \sigma_{S1}\rho_{11}\sigma_{T0}\rho_T \\ \rho_{10}\sigma_{T0}\sigma_{S1} - \sigma_{T0}\rho_T\sigma_{S1}\rho_{11} & \sigma_{T0}^2 - \sigma_{T0}^2\rho_T^2 \end{pmatrix}.$$

Then we solve $\rho_{10}\sigma_{S1}\sigma_{T0} - \sigma_{S1}\rho_{11}\sigma_{T0}\rho_T = 0$. The same process holds for the conditional model for $\theta_{10} = \theta_T \times \theta_{11}$.

Marginalization: Note the integral in equation 3 can be replaced with summation over the support of X for discrete covariates. Let $F_n(x)$ be the empirical distribution function of x, so for each value of s, equation 3 can be approximated by

$$\int_X \frac{(\gamma_{0,C} + \gamma_{1,C}s)\frac{1}{\epsilon_{S1}\sqrt{2\pi}} \exp(-\frac{1}{2\epsilon_{S1}^2}(s - (1\ x)^T(\omega_1\ \omega_2)^2)dF_n(x)}{f(s)}$$

Let $X_k, k = 1, ..., n$ denote the discrete values of X and let γ_{0,C_k} be the value of $\gamma_{0,C}$ at X_k. Then for a fixed s, we calculate $w_k = \exp(-\frac{1}{2\hat{\sigma}_{S1}^2}(s - (1\ X_k)^T(\hat{\omega_1}\ \hat{\omega_2}))^2)$ for each value of X_k, rescale w_k such that $\Sigma_k w_k = 1$, and calculate $\Sigma_x(\gamma_{0,C_k} + \gamma_{1,C}s)w_k$. This can be solved in closed form when X is normally distributed.

APPENDIX B

The Four Trivariate Normal Distributions and Corresponding Surrogacy Quantities

These distributions in the main text are derived from the four-dimensional, joint normal distribution

$$\begin{pmatrix} S(1) \\ T(0) \\ T(1) \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix}, \begin{pmatrix} \sigma_{S1}^2 & \sigma_{S1}\sigma_{T0}\rho_{10} & \sigma_{S1}\sigma_{T1}\rho_{11} & \sigma_{S1}\sigma_X\rho_{1X} \\ & \sigma_{T0}^2 & \sigma_{T0}\sigma_{T1}\rho_T & \sigma_{T0}\sigma_X\rho_{X0} \\ & & \sigma_{T1}^2 & \sigma_{T1}\sigma_X\rho_{X1} \\ & & & \sigma_X^2 \end{pmatrix}\right)$$

1. $$\begin{pmatrix} S(1) \\ T(0) \\ T(1) \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}, \begin{pmatrix} \sigma_{S1}^2 & \rho_{10}\sigma_{S1}\sigma_{T0} & \rho_{11}\sigma_{S1}\sigma_{T1} \\ & \sigma_{T0}^2 & \rho_T\sigma_{T0}\sigma_{T1} \\ & & \sigma_{T1}^2 \end{pmatrix}\right)$$

2. $$\left(\begin{array}{c|c} S(1) & X \\ T(0) & \\ T(1) & \end{array}\right) \sim N\left(\begin{pmatrix} \omega_1 + \omega_2 X \\ \omega_3 + \omega_4 X \\ \omega_5 + \omega_6 X \end{pmatrix}, \begin{pmatrix} \epsilon_{S1}^2 & \epsilon_{S1}\epsilon_{T0}\theta_{10} & \epsilon_{S1}\epsilon_{T1}\theta_{11} \\ & \epsilon_{T0}^2 & \epsilon_{T0}\epsilon_{T1}\theta_T \\ & & \epsilon_3^2 \end{pmatrix}\right)$$

$\gamma_{1,OC} = \frac{\epsilon_3\theta_{11} - \epsilon_2\theta_{10}}{\epsilon_1}$, $\gamma_{0,OC} = (\omega_5 + \omega_6 X - \omega_3 - \omega_4 X) - \gamma_{1,OC}(\omega_1 + \omega_2 X)$

$\gamma_{1,O} = \frac{\sigma_{T1}\rho_{11} - \sigma_{T0}\rho_{10}}{\sigma_{S1}}$, $\gamma_{0,O} = (\delta_3 - \delta_2) - \gamma_{1,O}\delta_1$

3. $$\begin{pmatrix} S(1) \\ T^D(0) \\ T^D(1) \end{pmatrix} \sim N\left(\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}, \begin{pmatrix} \tau_{S1}^2 & \tau_{S1}\tau_{T0}\pi_{10} & \tau_{S1}\tau_{T1}\pi_{11} \\ & \tau_{T0}^2 & \tau_{T0}\tau_{T1}\pi_T \\ & & \tau_{T1}^2 \end{pmatrix}\right)$$

4. $$\left(\begin{array}{c|c} S(1) & X \\ T^D(0) & \\ T^D(1) & \end{array}\right) \sim N\left(\begin{pmatrix} \phi_1 + \phi_2 X \\ \phi_3 + \phi_4 X \\ \phi_5 + \phi_6 X \end{pmatrix}, \begin{pmatrix} \xi_{S1}^2 & \xi_{S1}\xi_{T0}\psi_{10} & \xi_{S1}\xi_{T1}\psi_{11} \\ & \xi_{T0}^2 & \xi_{T0}\xi_{T1}\psi_T \\ & & \xi_{T1}^2 \end{pmatrix}\right)$$

$\gamma_{1,DC} = \frac{\xi_{T1}\psi_{11} - \xi_{T0}\psi_{10}}{\xi_{S1}}$, $\gamma_{0,DC} = (\phi_5 + \phi_6 X - \phi_3 - \phi_4 X) - \gamma_{1,DC}(\phi_1 + \phi_2 X)$

$\gamma_{1,D} = \frac{\tau_3\pi_{11} - \tau_2\pi_{10}}{\tau_1}$, $\gamma_{0,D} = (\eta_3 - \eta_2) - \gamma_{1,D}\eta_1$

The conditional quantities $\gamma_{1,OC}$ and $\gamma_{0,OC}$ can be calculated only from model 2, which can be subsequently marginalized over. The marginal quantities $\gamma_{1,O}$ and $\gamma_{0,O}$ are the same for models 1 and 2. Similarly, $\gamma_{1,DC}$ and $\gamma_{0,DC}$ from model 4 can be marginalized over to calculate the same $\gamma_{1,D}$ and $\gamma_{0,D}$ as in model 3.

APPENDIX C

Tables for Simulation Results Demonstrating Different Definitions of the Endpoint and Different Generating Parameter Values

Setting	Fit Conditional Independence Assumption	$\gamma_{0,M}$ Est	Bias	SE	SD	$\gamma_{1,M}$ Est	Bias	SE	SD	True Distance from Cond. Ind. on fit scale
3A	$T^D(0) \perp S(1)\|T^D(1)$	0.057	0.057	0.398	0.256	0.520	-0.030	0.175	0.093	0.002
3A	None	0.067	0.127	0.956	0.255	0.515	-0.065	0.468	0.093	0.002
4A	$T^D(0) \perp S(1)\|T(1), X$	0.058	0.118	0.422	0.478	0.527	-0.053	0.177	0.088	0.000
4A	None	0.091	0.151	0.934	0.483	0.519	-0.061	0.455	0.090	0.000
1A	$T(0) \perp S(1)\|T(1)$	0.068	0.068	0.414	0.303	0.516	-0.034	0.181	0.111	-0.098
1A	None	0.078	0.078	1.122	0.301	0.512	-0.038	0.552	0.110	-0.098
2A	$T(0) \perp S(1)\|T(1), X$	0.044	0.044	0.426	0.308	0.525	-0.025	0.179	0.091	0.000
2A	None	0.080	0.080	0.989	0.493	0.522	-0.028	0.478	0.092	0.000
3B	$T^D(0) \perp S(1)\|T^D(1)$	0.056	0.056	0.413	0.293	0.518	-0.032	0.181	0.108	-0.101
3B	None	0.066	0.066	1.117	0.291	0.514	-0.036	0.549	0.107	-0.101
4B	$T^D(0) \perp S(1)\|T^D(1), X$	0.052	0.052	0.424	0.479	0.522	-0.028	0.180	0.084	0.000
4B	None	0.099	0.099	0.978	0.480	0.501	-0.049	0.471	0.085	0.000
1B	$T(0) \perp S(1)\|T(1)$	0.058	0.058	0.398	0.256	0.519	-0.031	0.175	0.093	0.000
1B	None	0.067	0.067	0.955	0.255	0.516	-0.034	0.468	0.093	0.000
2B	$T(0) \perp S(1)\|T(1), X$	0.058	0.058	0.422	0.478	0.527	-0.023	0.177	0.088	0.000
2B	None	0.082	0.082	0.998	0.479	0.517	-0.033	0.477	0.088	0.000
3C	$T^D(0) \perp S(1)\|T^D(1)$	-0.974	0.046	0.399	0.286	0.534	-0.026	0.172	0.110	0.003
3C	None	-0.963	0.057	1.015	0.284	0.530	-0.030	0.497	0.109	0.003
4C	$T^D(0) \perp S(1)\|T^D(1), X$	-0.944	0.076	0.422	0.478	0.529	-0.031	0.177	0.115	0.000
4C	None	-0.904	0.116	1.002	0.483	0.506	-0.054	0.478	0.109	0.000
1C	$T(0) \perp S(1)\|T(1)$	-0.905	0.095	0.418	0.266	0.501	-0.049	0.188	0.093	0.003
1C	None	-0.895	0.105	1.056	0.265	0.497	-0.053	0.520	0.092	0.003
2C	$T(0) \perp S(1)\|T(1), X$	-0.944	0.056	0.422	0.478	0.529	-0.021	0.177	0.115	0.000
2C	None	-0.919	0.081	0.998	0.479	0.519	-0.031	0.477	0.116	0.000
3D	$T^D(0) \perp S(1)\|T^D(1)$	-1.288	0.045	0.324	0.304	0.191	-0.029	0.130	0.094	0.000
3D	None	-1.288	0.045	1.567	0.301	0.193	-0.027	0.779	0.092	0.000
4D	$T^D(0) \perp S(1)\|T^D(1), X$	-1.370	-0.037	0.292	0.513	0.231	0.011	0.103	0.136	0.000
4D	None	-1.321	0.012	1.145	0.513	0.209	-0.011	0.555	0.136	0.000
1D	$T(0) \perp S(1)\|T(1)$	-1.257	0.093	0.353	0.360	0.177	-0.043	0.145	0.100	-0.094
1D	None	-1.261	0.089	1.996	0.357	0.181	-0.039	0.995	0.097	-0.094
2D	$T(0) \perp S(1)\|T(1), X$	-1.371	-0.021	0.291	0.512	0.232	0.012	0.099	0.189	0.000
2D	None	-1.343	0.007	1.167	0.511	0.222	0.002	0.563	0.189	0.000
3E	$T^D(0) \perp S(1)\|T^D(1)$	1.009	0.059	0.210	0.207	0.181	-0.129	0.116	0.105	0.415
3E	None	1.011	0.061	0.359	0.207	0.182	-0.128	0.564	0.102	0.415
4E	$T^D(0) \perp S(1)\|T^D(1), X$	1.011	0.061	0.212	0.466	0.178	-0.132	0.118	0.106	0.414
4E	None	1.002	0.052	0.364	0.466	0.195	-0.115	0.565	0.104	0.414
1E	$T(0) \perp S(1)\|T(1)$	1.012	0.063	0.219	0.233	0.181	-0.119	0.128	0.118	0.414
1E	None	1.014	0.065	0.392	0.232	0.183	-0.117	0.630	0.115	0.414
2E	$T(0) \perp S(1)\|T(1), X$	1.011	0.062	0.212	0.466	0.178	-0.122	0.118	0.106	0.414
2E	None	1.002	0.053	0.364	0.466	0.195	-0.105	0.565	0.104	0.414
1F	$T(0) \perp S(1)\|T(1)$	1.360	0.050	0.431	0.376	0.553	-0.027	0.187	0.148	0.183
1F	None	1.387	0.077	1.163	0.375	0.539	-0.041	0.568	0.148	0.183
2F	$T(0) \perp S(1)\|T(1), X$	1.485	0.175	0.402	0.323	0.501	-0.079	0.170	0.222	0.000
2F	None	1.501	0.191	0.977	0.322	0.498	-0.082	0.475	0.218	0.000

Table 3.1: Simulation results demonstrating different definitions of the endpoint (settings 1-4) and different generating parameter values ($A - F$). Estimates and bias are shown here since true generating values differ across scenarios. These are shown in the main text Table 2.1.

APPENDIX D

Simulation Results and Sensitivity Analysis of Data Example Results

Figure 4.1: Simulation results and sensitivity analysis of data example results over different values of θ_T using the observed data method and conditional independence assumption. The models are fit by either fixing the value of θ_T or placing the prior distribution over the parameter that we consider in the main text.

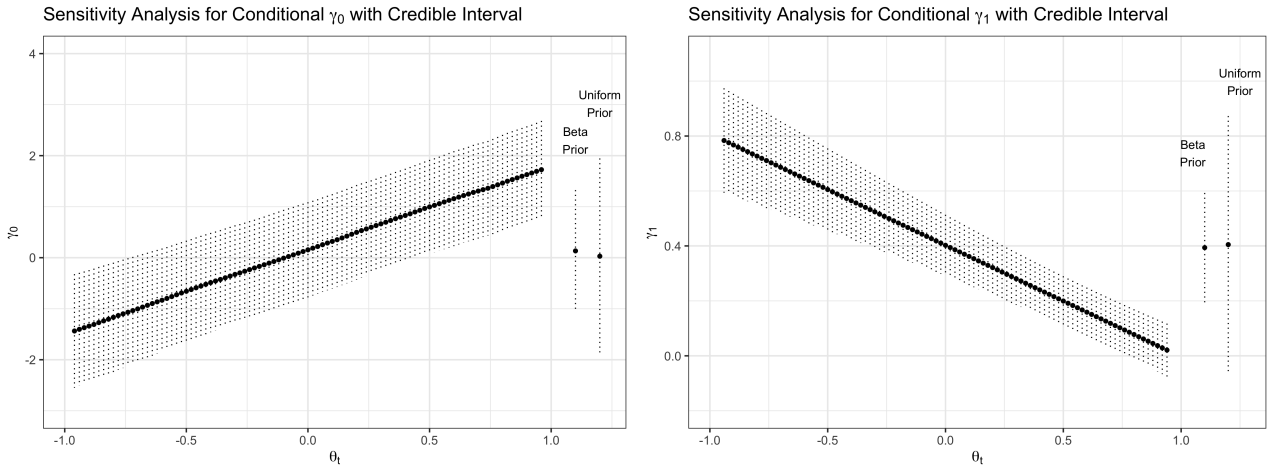
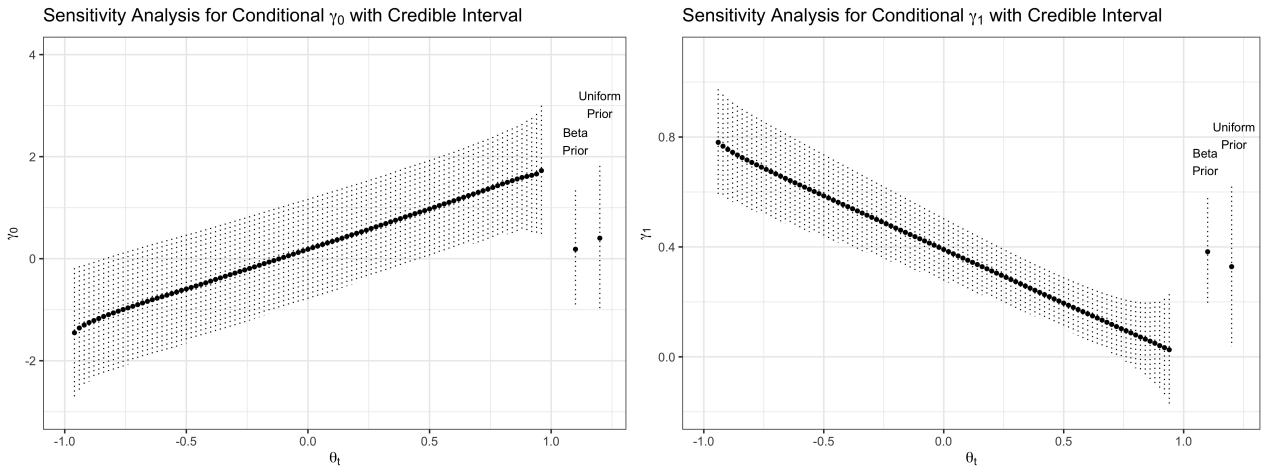


Figure 4.2: Simulation results and sensitivity analysis of data example results over different values of θ_T using the imputation method and conditional independence assumption.



The actual priors we used are Beta(2.7, 5) and Uniform(-1, 1). This is compared to fixing θ_T at possible values one at a time and repeating the process. These plots show that we would say $\gamma_0 > 0$ if $\theta_T > 0.48$, and $\gamma_1 > 0$ if $\theta_T < 0.75$ approximately. Also, we see that both the observed data only and imputation methods return nearly the same results (perhaps except at the tails where convergence at the boundaries of the parameter space may play a role).

APPENDIX E

Simulation Results Demonstrating Effect of Non-normal Distributions

Setting	Fit Conditional Independence Assumption	True Distribution	$\gamma_{0,M}$ Estimate	$\gamma_{0,M}$ Bias	$\gamma_{0,M}$ SE	$\gamma_{0,M}$ SD	$\gamma_{1,M}$ Estimate	$\gamma_{1,M}$ Bias	$\gamma_{1,M}$ SE	$\gamma_{1,M}$ SD
2A	$T(0) \perp S(1) \mid T(1), X$	T	0.067	0.127	0.404	0.308	0.512	-0.068	0.171	0.124
2B	$T(0) \perp S(1) \mid T(1), X$	T	0.067	0.067	0.404	0.308	0.512	-0.038	0.171	0.124
2C	$T(0) \perp S(1) \mid T(1), X$	T	-0.937	0.063	0.402	0.312	0.513	-0.037	0.170	0.123
2D	$T(0) \perp S(1) \mid T(1), X$	T	-1.371	-0.021	0.285	0.378	0.229	0.009	0.097	0.150
2E	$T(0) \perp S(1) \mid T(1), X$	T	0.994	0.044	0.212	0.218	0.192	-0.118	0.090	0.153
2F	$T(0) \perp S(1) \mid T(1), X$	T	1.452	0.140	0.416	0.339	0.508	-0.068	0.174	0.118
2A	$T(0) \perp S(1) \mid T(1), X$	Gamma	0.133	0.193	0.396	0.252	0.481	-0.099	0.165	0.138
2B	$T(0) \perp S(1) \mid T(1), X$	Gamma	0.133	0.133	0.397	0.252	0.485	-0.065	0.166	0.137
2C	$T(0) \perp S(1) \mid T(1), X$	Gamma	-0.868	0.132	0.396	0.258	0.481	-0.069	0.165	0.138
2D	$T(0) \perp S(1) \mid T(1), X$	Gamma	-1.311	0.039	0.280	0.308	0.200	-0.020	0.093	0.128
2E	$T(0) \perp S(1) \mid T(1), X$	Gamma	1.015	0.065	0.210	0.194	0.164	-0.136	0.084	0.127
2F	$T(0) \perp S(1) \mid T(1), X$	Gamma	1.475	0.163	0.401	0.306	0.488	-0.088	0.168	0.149

Table 5.1: Simulation results demonstrating effect of non-normal distributions when fitting conditional models that assume normality. Results are shown on the original scale and can be compared to other supplemental tables.

APPENDIX F

Generative Parameter Values for Plausible Clinical Trial Data Example

We generate the $S(1), T(0), T(1)$ outcomes at one time point and effects of age, A, and baseline, X.

$$\begin{pmatrix} X \\ A \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 \\ & \sigma_2^2 \end{pmatrix}\right)$$

We chose values for the means and variances: $\delta_1 = 24, \delta_2 = 5$, $\sigma_1^2 = 1, \sigma_2^2 = 0.65$, $\rho_1 = 0.7$. We expect there to be a quadratic effect of age, so we generate

$$\left(\begin{array}{c|c} S(1) & X, \\ T(0) & A, \\ T(1) & A^2 \end{array}\right) \sim N\left(\begin{pmatrix} \phi_1 + \phi_2 X + \phi_3 A \\ \phi_4 + \phi_5 X + \phi_6 A + \phi_7 A^2 \\ \phi_8 + \phi_9 X + \phi_{10} A + \phi_{11} A^2 \end{pmatrix}, \begin{pmatrix} \xi_{S1}^2 & \xi_{S1}\xi_{T0}\psi_{10} & \xi_{S1}\xi_{T1}\psi_{11} \\ & \xi_{T0}^2 & \xi_{T0}\xi_{T1}\psi_T \\ & & \xi_{T1}^2 \end{pmatrix}\right)$$

$\phi_1 = 3.8, \phi_2 = 0, \phi_3 = 0$, $\phi_4 = 0.1, \phi_5 = 1.14, \phi_6 = 0.45, \phi_7 = -0.2$, $\phi_8 = 10.6, \phi_9 = 1.1$, $\phi_{10} = -1.15, \phi_{11} = -0.2$, $\xi_{S1}^2 = 1, \xi_{T0}^2 = 0.35, \xi_{T1}^2 = 0.35$, $\psi_{10} = 0.013, \psi_{11} = 0.65, \psi_T = 0.02$

APPENDIX G

Counterfactual Imputation Details

Here we look at the form of the covariance for $(S(1), T(0), T(1), b^{(0)}, b^{(1)})$ for the random intercept model with no effect of time. Here we assume each repeated measurement in $\boldsymbol{T(z)}$ has the same distribution, so for any specific j

$$\begin{pmatrix} S(1)_i & T(0)_{ij} & T(1)_{ij} & b_i^{(0)} & b_i^{(1)} \end{pmatrix}^T \sim MVN\left(\begin{pmatrix} \alpha_1 & \beta_0 & \beta_1 & 0 & 0 \end{pmatrix}^T, \Psi\right)$$

if fixed and random effects, respectively denoted X and Z, are scalars and $\boldsymbol{\beta}_0 = \beta_0$ and $\boldsymbol{\beta}_1 = \beta_1$.

$$\Psi = \begin{pmatrix} \sigma^2_{S1} & \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{11}\sigma_{S1}\sigma_{b_1} & \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ & \sigma^2_{b_0} + \sigma^2_e & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma^2_{b_0} & \rho_T\sigma_{b_0}\sigma_{b_1} \\ & & \sigma^2_{b_1} + \sigma^2_e & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma^2_{b_1} \\ & & & \sigma^2_{b_0} & \rho_T\sigma_{b_0}\sigma_{b_1} \\ & & & & \sigma^2_{b_1} \end{pmatrix}$$

While this is written in condensed form here, the vectors $\boldsymbol{T}(z)$ will result in a larger block structured covariance matrix Ψ of dimension $2m + 3$.

$$\begin{pmatrix}
\sigma_{S1}^2 & \rho_{10}\sigma_{S1}\sigma_{b_0} & \cdots & \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{11}\sigma_{S1}\sigma_{b_1} & \cdots & \rho_{11}\sigma_{S1}\sigma_{b_1} & \rho_{11}\sigma_{S1}\sigma_{b_1} & \rho_{10}\sigma_{S1}\sigma_{b_0} \\
 & \sigma_{b_0}^2+\sigma_e^2 & \cdots & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} & \cdots & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\
 & & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & & & \sigma_{b_0}^2+\sigma_e^2 & \rho_T\sigma_{b_0}\sigma_{b_1} & \cdots & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\
 & & & & \sigma_{b_1}^2+\sigma_e^2 & \cdots & \sigma_{b_1}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \\
 & & & & & \ddots & \vdots & \vdots & \vdots \\
 & & & & & & \sigma_{b_1}^2+\sigma_e^2 & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \\
 & & & & & & & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\
 & & & & & & & & \sigma_{b_1}^2
\end{pmatrix}$$

In general, the conditional distribution of $x_1|x_2$ follows $N(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(a-\mu_2), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

We impute the counterfactual values $S(1), \boldsymbol{T(0)}, \boldsymbol{T(1)}$ from the distributions of

$$\begin{pmatrix} \boldsymbol{T(0)} \\ \hline S(1) \end{pmatrix} \begin{vmatrix} b^{(0)}, \boldsymbol{T(1)}, b^{(1)} \\ \hline \mu, \rho, \sigma \end{vmatrix} \sim N(\mu^{'}, \Sigma^{'}), \mu^{'} = \mu_1^{'} + \Sigma_{12}^{'}\Sigma_{22}^{-1'}(a^{'} - \mu_2^{'}), \Sigma^{'} = \Sigma_{11}^{'} - \Sigma_{12}^{'}\Sigma_{22}^{-1'}\Sigma_{21}^{'}$$ $\mu_1^{'} = \beta_0, a^{'} = (b^{(0)} \quad \boldsymbol{T(1)} \quad b^{(1)} \quad S(1)), \mu_2^{'} = (0 \quad \beta_1 \quad 0 \quad \alpha_1), \Sigma_{12}^{'} = \begin{pmatrix} \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \end{pmatrix}$,

$$\Sigma_{11}^{'} = \sigma_{b_0}^2 + \sigma_e^2, \Sigma_{22}^{'} = \begin{pmatrix} \sigma_{S1}^2 & \rho_{11}\sigma_{S1}\sigma_{b_1} & \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ & \sigma_{b_1}^2+\sigma_e^2 & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \\ & & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\ & & & \sigma_{b_1}^2 \end{pmatrix}$$

$$\begin{pmatrix} S(1) \\ \boldsymbol{T(1)} \end{pmatrix} \begin{vmatrix} b^{(0)}, \boldsymbol{T(0)}, \\ b^{(1)}, \mu, \rho, \sigma \end{vmatrix} \sim N(\mu^{''}, \Sigma^{''}), \mu^{''} = \mu_1^{''} + \Sigma_{12}^{''}\Sigma_{22}^{-1''}(a^{''} - \mu_2^{''}), \Sigma^{''} = \Sigma_{11}^{''} - \Sigma_{12}^{''}\Sigma_{22}^{-1''}\Sigma_{21}^{''}$$ $\mu_1^{''} = (\alpha_1 \quad \beta_1), a^{''} = (b^{(0)} \quad \boldsymbol{T(0)} \quad b^{(1)}), \mu_2^{''} = (0 \quad \beta_1 \quad 0), \Sigma_{12}^{''} = \begin{pmatrix} \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{10}\sigma_{S1}\sigma_{b_0} & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ \rho_T\sigma_{b_0}\sigma_{b_1} & \rho_T\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$,

$$\Sigma_{11}^{'} = \begin{pmatrix} \sigma_{S1}^2 & \rho_{11}\sigma_{S1}\sigma_{b_1} \\ & \sigma_{b_1}^2+\sigma_e^2 \end{pmatrix}, \Sigma_{22}^{'} = \begin{pmatrix} \sigma_{b_0}^2+\sigma_e^2 & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\ & \sigma_{b_0}^2 & \rho_T\sigma_{b_0}\sigma_{b_1} \\ & & \sigma_{b_1}^2 \end{pmatrix}$$

For each, the conditional mean and covariances are calculated from the Ψ matrix. We draw the

random effect estimates using related normal distributions with the conditioning terms changed. This matrix will depend on time as the models become more complex. More details are shown in the following section and further derivation for similar models is given in Schafer and Yucel (2002).

APPENDIX H

Derivations for Random Slopes Details

For shorthand, let $\boldsymbol{\Psi} = \begin{pmatrix} \sigma^2_{S1} & \sigma_{S101} & \sigma_{S111} & \sigma_{S102} & \sigma_{S112} \\ & \sigma^2_{b01} & \sigma_{0111} & \sigma_{0102} & \sigma_{0112} \\ & & \sigma^2_{b11} & \sigma_{1102} & \sigma_{1112} \\ & & & \sigma^2_{b02} & \sigma_{0212} \\ & & & & \sigma^2_{b12} \end{pmatrix}$

T_{BL} is a measure of $T(0)$ at time 0 and is equal to $\beta_0^{(0)} + b_0^{(0)} + e$. The two models treating T_{BL} as either an outcome or a baseline covariate can be equated algebraically after integrating over the random effects. Specifically, if we treat T_{BL_i} as $T(0)_{i0}$, we have:

$$\begin{pmatrix} S(1)_i & \boldsymbol{T(0)}_i & \boldsymbol{T(1)}_i & T_{BL_i} \end{pmatrix} \sim MVN(\begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_0^{(0)} + \boldsymbol{\beta}_2^{(0)} \boldsymbol{t}_i \\ \boldsymbol{\beta}_0^{(1)} + \boldsymbol{\beta}_2^{(1)} \boldsymbol{t}_i \\ \beta_0^{(0)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ & \Sigma_{22} \end{pmatrix}) \qquad (8.1)$$

where we have parameterized the covariance matrix in this block structure to use properties of conditional normal distributions. Σ_{22} is the variance of T_{BL}, Σ_{11} is the covariance matrix of $S(1), \boldsymbol{T(0)}$, and $\boldsymbol{T(1)}$, and Σ_{12} is the covariance between these.

When conditioning on T_{BL_i} as a baseline covariate, we obtain:

$$\left(\begin{array}{c|c} S(1)_i & \\ \boldsymbol{T(0)}_i, & T_{BL_i} \\ \boldsymbol{T(1)}_i & \end{array}\right) \sim MVN(\left(\begin{array}{c} \alpha_1^{'} + \alpha_2^{'} T_{BL_i} \\ \boldsymbol{\beta}_0^{(0)'} + \boldsymbol{\beta}_1^{(0)'} T_{BL_i} + \boldsymbol{\beta}_2^{(0)'} \boldsymbol{t}_i \\ \boldsymbol{\beta}_0^{(1)'} + \boldsymbol{\beta}_1^{(1)'} T_{BL_i} + \boldsymbol{\beta}_2^{(1)'} \boldsymbol{t}_i \end{array}\right), \Omega) \tag{8.2}$$

where $\Omega = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ denotes the covariance of the conditional outcomes. The vector of coefficients, denoted by $'$ and corresponding to T_{BL_i} for each potential outcome, is also an algebraic function of Σ_{12} and Σ_{22}. If we chose to model T_{BL} as a covariate, after fitting the model in Eq. 2, we will have to integrate over T_{BL} to obtain marginal quantities of our validation metrics.

We also note for $j = 1, ..., m$, $cov(T_{BL}, S(1)) = cov(b_0^{(0)} + e, S(1)) = \sigma_{S101}$, $cov(T_{BL}, T(0)_j) = cov(b_0^{(0)}, b_0^{(0)} + b_1^{(0)} j) = \sigma_{b01}^2 + j\sigma_{0111}$, $cov(T_{BL}, T(1)_j) = cov(b_0^{(0)}, b_0^{(1)} + b_1^{(1)} j) = \sigma_{0102} + j\sigma_{0112}$. Let $\Sigma_{11} =$

$$\left(\begin{array}{ccccccc}
\sigma_{S1}^2 & \sigma_{S101} + \sigma_{S111} & \dots & \sigma_{S101} + m\sigma_{S111} & \sigma_{S102} + \sigma_{S1}\sigma_{b12}\rho_{S112} & \dots & \sigma_{S102} + m\sigma_{S1}\sigma_{b12}\rho_{S112} \\
 & \sigma_{b01}^2 + \sigma_{b11}^2 + \sigma_e^2 + \sigma_{0111} + \sigma_{0111} & \dots & \sigma_{b01}^2 + \sigma_{0111} + m\sigma_{0111} + m\sigma_{b11}^2 & \sigma_{0102} + \sigma_{0112} + \sigma_{1112} + \sigma_{1102} & \dots & \sigma_{0102} + m\sigma_{0112} + m\sigma_{1112} + \sigma_{1102} \\
 & & \ddots & \vdots & \vdots & \vdots & \vdots \\
 & & & \sigma_{b01}^2 + m^2\sigma_{b11}^2 + \sigma_e^2 + \sigma_{0111} + m\sigma_{0111} & \sigma_{0102} + m\sigma_{0112} + m\sigma_{1112} + m\sigma_{1102} & \dots & \sigma_{0102} + m\sigma_{0112} + m\sigma_{1112} + m\sigma_{1102} \\
 & & & & \sigma_{b02}^2 + \sigma_{b12}^2 + \sigma_e^2 + \sigma_{0212} + \sigma_{0212} & \dots & \sigma_{b02}^2 + m\sigma_{b12}^2 + \sigma_{0212} + m\sigma_{0212} \\
 & & & & & \ddots & \vdots \\
 & & & & & & \sigma_{b02}^2 + m^2\sigma_{b12}^2 + m\sigma_{0212} + m\sigma_{0212} + \sigma_e^2
\end{array}\right)$$

$$\Sigma_{22} = \left(\begin{array}{c} \sigma_{b0}^2 + \sigma_e^2 \end{array}\right), \Sigma_{12} = \left(\begin{array}{cccccc} \sigma_{S101} & \sigma_{b01}^2 + \sigma_{0111} & \dots & \sigma_{b01}^2 + m\sigma_{0111} & \sigma_{0102} + \sigma_{0112} & \dots & \sigma_{0102} + m\sigma_{0112} \end{array}\right)$$

When considering the coefficient for T_{BL_i} in Eq. 2, and ultimately what the γ quantities will look like, the coefficients for T_{BL} will not be equal between the $T(0)$ and $T(1)$ outcomes unless $\sigma_{0102} + t\sigma_{0112} = \sigma_{b01}^2 + t\sigma_{0111}$. Since there is an interaction between time and the effect of T_{BL}, rather than considering time-varying coefficients, we write the second and third terms as $\frac{\sigma_{b01}^2}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)}) + \frac{\sigma_{b11}\sigma_{b01}\rho_{0111}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)})\boldsymbol{t}_i$, $\frac{\sigma_{b02}\sigma_{b01}\rho_{0102}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)}) + \frac{\sigma_{b01}\sigma_{b12}\rho_{0112}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)})\boldsymbol{t}_i$. From this, $E(T(1) - T(0)|s) = \beta_0^{(1)} + \beta_2^{(1)} t_{ij} + \frac{\sigma_{b02}\sigma_{b01}\rho_{0102}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)}) + \frac{\sigma_{b01}\sigma_{b12}\rho_{0112}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)})t_{ij} - \beta_0^{(0)} - \beta_2^{(0)} t_{ij} - (\frac{\sigma_{b01}^2}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)}) + \frac{\sigma_{b11}\sigma_{b01}\rho_{0111}}{\sigma_{b01}^2+\sigma_e^2}(T_{BL} - \beta_0^{(0)})t_{ij}) + \gamma_1 s = \gamma_0^*(t_{ij}) + \gamma_1^*(t_{ij})s$.

In order to integrate over T_{BL}, we need to calculate the conditional form of $\gamma_0^*(t_{ij})$ and $\gamma_1^*(t_{ij})$. When integrating over a baseline covariate B,

$$\int_B E(T(1) - T(0)|S(1), B = b) f(B|S(1) = s) db = \int_B (\gamma_0^* + \gamma_1^* s) f(B|S(1) = s) db$$

We will calculate the empirical quantity of this marginal expectation for a fixed time t_{ij} using Bayes rule. Let the subscript C denote the conditional quantities for a given set of covariates and k be an index to denote the value X for a given s. Our empirical integration is calculated as

$$\sum_X \frac{(\gamma_{0,C_k} + \gamma_{1,C} s) \frac{1}{|\Omega|\sqrt{2\pi}} \exp(-\frac{1}{2}(s - (\boldsymbol{X}_k)^T(\boldsymbol{\Phi}))\Omega^{-1}(s - (\boldsymbol{X}_k)^T(\boldsymbol{\Phi})) dF_n(x)}{f(s)}$$

where X is continuous and $dF_n(x)$ is an empirical distribution, summed for each particular value in $S(1) = s$. The conditional distribution of $S(1)|\boldsymbol{X}$ will follow a Normal distribution. For a fixed s, we calculate $w_k = \exp(-\frac{1}{2}(s - (\boldsymbol{X}_k)^T(\hat{\boldsymbol{\Phi}}))\Omega^{-1}(s - (\boldsymbol{X}_k)^T(\hat{\boldsymbol{\Phi}}))$ for each value of X_k, rescale w_k such that $\Sigma_k w_k = 1$, and calculate $\Sigma_x(\gamma_0^* + \gamma_1^* s) w_k$. Once we solve for these quantities, we summarize the marginal effect over values of s by fitting a linear model to get γ_0 and γ_1.

APPENDIX I

Fisher Approximation for Correlation Details

We use the Fisher z-transformation as an approximation for the posterior distribution of a correlation parameter ρ and use this approximation as the proposal distribution in the Metropolis-Hastings step within the MCMC algorithm. Fisher's z transformation from $\rho \rightarrow z$, also known as the arctanh, can be written $h(\rho) = 0.5\log((1+\rho)/(1-\rho))$ with derivative $h(\rho)' = 1/(1-\rho^2)$. We follow the steps:

1. Calculate $\hat{\rho}$ from the observed data or imputed counterfactual data as the empirical correlation between the random effects (for ρ_{11} it is the correlation between S and $b^{(1)}$). Let the previous value of ρ be ρ_{-1}. Since we expect ρ_{-1} and $\hat{\rho}$ to be similar, we will only consider ρ_{-1}.

2. Draw a proposed, transformed value of ρ where $z = h(\rho)$: $z \sim Normal(0.5\log((1+\rho_{-1})/(1-\rho_{-1})), 1/(n-3))$

3. Let $\rho^{'} = h^{-1}(z)$.

4. Because of the transformation, the density of ρ includes a Jacobian term. This is written $g(\rho|\rho_{-1}) = f(z|\rho_{-1})(\partial z/\partial\rho) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-1/2\sigma^2)(z(\rho) - z(\rho_{-1}))^2 \times \frac{1}{1-\rho^2}$, where $\sigma^2 = \frac{1}{n-3}$.

5. Calculate $g(\rho_{-1}|\rho) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-1/2\sigma^2)(z(\rho_{-1}) - z(\rho))^2 \times \frac{1}{1-\rho_{-1}{}^2}$.

6. Using the previous ρ_{-1}, the exponential terms will cancel when calculating $g(\rho^{'}|\rho_{-1})/g(\rho_{-1}|\rho^{'})$, leaving $\frac{1}{1-\rho_{-1}^2}/\frac{1}{1-\rho^{'2}}$ from the ratio of the values from 4 and 5.

7. The new value $\rho^{'} = h^{-1}(z)$ is accepted with probability $min(1, \frac{P(\rho^{'})}{P(\rho_{-1})}\frac{1}{1-\rho_{-1}^2}/\frac{1}{1-\rho^{'2}})$. For ρ_T, we would like to incorporate an informative prior distribution, so we evaluate the posterior P distribution for both the numerator and the denominator using the rescaled Beta(8, 5) distribution. P is the product of the likelihood and the prior and can be written as $\rho \propto |R|^{-n/2} \exp\left(-\frac{1}{2}\Sigma_{i=1}^{n}(b_{\cdot}^{(0)}\ b_{\cdot}^{(1)})\Psi^{-1}(b_{\cdot}^{(0)}\ b_{\cdot}^{(1)})^T\right)(1-\rho)^{(5-1)}(\rho+1)^{(8-1)}$ where $\Psi = SRS$ is decomposed into matrices containing standard deviations (S) and correlations (R).

   The MH step is computationally fast and can be run for many MCMC iterations (10,000 or more) in a few hours.

APPENDIX J

Delayed-Start Treatment Design Details

To develop the model in the Delayed-start treatment design, we first consider models for the natural history of the outcome for a patient who does not receive the intervention compared to that patient receiving the intervention at an early age. Let A_i denote current age and B_i age at baseline where $A_i = B_i + t_i$. Ignoring the measurement error term, $T_i(0, A_i) = \beta_0^{(0)} + \beta_1^{(0)} A_i + \beta_2 A_i^2 + b_{i0}^{(0)} + b_{i1}^{(0)} A_i$. Similarly $T_i(1, A_i) = \beta_0^{(1)} + \beta_1^{(1)} A_i + \beta_2 A_i^2 + b_{i0}^{(1)} + b_{i1}^{(1)} A_i$. Let V be time on study of the delayed-treatment arm.

Since we model an individual's repeated measurements while they receive the treatment, consider what the smooth trajectory of outcomes would look like for an individual in the delayed-treatment group. At time of intervention, the immediate change in the fixed effect is $\beta_0^{(1)} - \beta_0^{(0)}$, and there is a change in the slope for age where the new fixed effect slope is $\beta_1^{(1)}$ at the time of intervention. The random intercept changes from $b_{i0}^{(0)}$ to $b_{i0}^{(1)}$, and the random slope from that time forward is $b_{i1}^{(1)}$. Then for individuals in the $z = 0$ arm, their value of T prior to any crossover is given by $T_i(0, t_i, B_i) = \beta_0^{(0)} + \beta_1^{(0)}(t_i + B_i) + \beta_2(t_i + B_i)^2 + b_{0i}^{(0)} + b_{1i}^{(0)}(t_i + B_i)$. For an individual who starts the study at age B_i, their baseline value is given by $T_i(0, 0, B_i) = \beta_0^{(0)} + \beta_1^{(0)} B_i + \beta_2 B_i^2 + b_{0i}^{(0)} + b_{1i}^{(0)} B_i$
For those randomized to $z = 1$, their value right after randomization is given by $T_i(1, 0^+, B_i) = \beta_0^{(1)} + \beta_1^{(0)} B_i + \beta_2 B_i^2 + b_{0i}^{(1)} + b_{1i}^{(0)} B_i$ and for $t > 0$, their values are $T_i(1, t, B_i) = \beta_0^{(1)} + \beta_1^{(0)} B_i + \beta_1^{(1)} t_i + \beta_2 (B_i + t)^2 + b_{0i}^{(1)} + b_{1i}^{(0)} B_i + b_{1i}^{(1)} t$
For those who crossover at age $\tau_i = B_i + V$, for $age < \tau_i$ an individual's trajectory follows the described $z = 0$ model. Just prior to V, their value of $T_i = \beta_0^{(0)} + \beta_1^{(0)} \tau_i + \beta_2^{(0)} \tau_i^2 + b_{0i}^{(0)} + b_{1i}^{(0)} \tau_i$

$= \beta_0^{(0)} + \beta_1^{(0)}(B_i + V) + \beta_2^{(0)}(B_i + V)^2 + b_{0i}^{(0)} + b_{1i}^{(0)}(B_i + V)$. At $t = V^+, T_i = \beta_0^{(1)} + \beta_1^{(0)}(B_i + V) + \beta_2(B_i + V)^2 + b_{0i}^{(1)} + b_{1i}^{(0)}(B_i + V)$, and for $t > V, T_i = \beta_0^{(1)} + \beta_1^{(0)}(B_i + V) + \beta_1^{(1)}(t - V) + \beta_2(B_i + V)^2 + b_{0i}^{(1)} + b_{1i}^{(0)}(B_i + V) + b_{1i}^{(1)}(t - V)$.

We assume the observed data model for $z = 0$ in a regular design with a baseline measurement is

$$\begin{pmatrix} 1 & t_0 + B_i & 0 & 0 & (t_0 + B_i)^2 \\ & & \vdots & & \\ 1 & t_j + B_i & 0 & 0 & (t_j + B_i)^2 \\ & & \vdots & & \\ 1 & t_m + B_i & 0 & 0 & (t_m + B_i)^2 \end{pmatrix} \begin{pmatrix} \beta_0^{(0)} \\ \beta_1^{(0)} \\ \beta_0^{(1)} \\ \beta_1^{(1)} \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & t_0 + B_i & 0 & 0 \\ & \vdots & & \\ 1 & t_j + B_i & 0 & 0 \\ & \vdots & & \\ 1 & t_m + B_i & 0 & 0 \end{pmatrix} \begin{pmatrix} b_{0i}^{(0)} \\ b_{1i}^{(0)} \\ b_{0i}^{(1)} \\ b_{1i}^{(1)} \end{pmatrix}$$

For $z = 1$,

$$\begin{pmatrix} 1 & B_i & 0 & 0 & B_i^2 \\ 0 & B_i & 1 & t_1 & (t_1 + B_i)^2 \\ & & \vdots & & \\ 0 & B_i & 1 & t_j & (t_j + B_i)^2 \\ & & \vdots & & \\ 0 & B_i & 1 & t_m & (t_m + B_i)^2 \end{pmatrix} \begin{pmatrix} \beta_0^{(0)} \\ \beta_1^{(0)} \\ \beta_0^{(1)} \\ \beta_1^{(1)} \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & B_i & 0 & 0 \\ 0 & B_i & 1 & t_1 \\ & \vdots & & \\ 0 & B_i & 1 & t_j \\ & \vdots & & \\ 0 & B_i & 1 & t_m \end{pmatrix} \begin{pmatrix} b_{0i}^{(0)} \\ b_{1i}^{(0)} \\ b_{0i}^{(1)} \\ b_{1i}^{(1)} \end{pmatrix}$$

For the delayed-treatment arm,

$$\begin{pmatrix} 1 & B_i & 0 & 0 & B_i^2 \\ 1 & t_1+B_i & 0 & 0 & (t_1+B_i)^2 \\ & & \vdots & & \\ 1 & t_j+B_i & 0 & 0 & (t_j+B_i)^2 \\ & & \vdots & & \\ 0 & V+B_i & 1 & 1 & (V+B_i)^2 \\ & & \vdots & & \\ 0 & V+B_i & 1 & t_m-V & (t_m+B_i)^2 \end{pmatrix} \begin{pmatrix} \beta_0^{(0)} \\ \beta_1^{(0)} \\ \beta_0^{(1)} \\ \beta_1^{(1)} \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & B_i & 0 & 0 \\ 1 & t_1+B_i & 0 & 0 \\ & \vdots & & \\ 1 & t_j+B_i & 0 & 0 \\ & \vdots & & \\ 0 & V+B_i & 1 & 1 \\ & \vdots & & \\ 0 & V+B_i & 1 & t_m-V \end{pmatrix} \begin{pmatrix} b_{0i}^{(0)} \\ b_{1i}^{(0)} \\ b_{0i}^{(1)} \\ b_{1i}^{(1)} \end{pmatrix}$$

APPENDIX K

Generating Proper Covariance Matrices

Here we provide more details about prior distributions and generating positive definite matrices. We have described using rescaled Beta$(3, 3)$ and Beta$(8, 5)$ priors in two separate scenarios. The means (standard deviations) of these distributions are 0 (0.378) and 0.231 (0.260) respectively. While it may seem a bit arbitrary to use these Beta distributions, they are motivated by empirical observations. The first symmetric Beta distribution is meant to be noninformative for the purpose of generating 'true' values in the infinite data case. We have looked extensively at the marginal distribution of the correlations after the positive definiteness rejections, and they tend to be unimodal (under the setting of no conditional independence assumptions). In our exploration of generating valid 5×5 covariance matrices with marginally Uniform priors on each correlation in the setting of random slopes models, we found the acceptance rate of the matrices to be so low that it was computationally prohibitive to draw thousands of valid covariance matrices when all ten correlation parameters are drawn marginally from a Uniform(-1, 1) distribution. To avoid this, we simulated from Beta distributions. We are deliberately using a Beta distribution to (largely) exclude correlations that are very close to -1 and +1. Further, we found that the distribution of valid draws of the correlations tended to be approximately Beta distributed after rejecting extreme correlation values that are unlikely to form a valid covariance matrix. We compared this method of generating matrices to the LKJ way of simulating (non informative) correlation matrices (Lewandowski, Kurowicka, and Joe, 2009) and found that the resulting marginal distributions of the correlation parameters were similar. We also established that empirically the draws of the ρ's were more or

less independent of each other even after the positive definiteness rejections. Separately from this, the second, non-symmetric Beta distribution is meant to be more informative for the purpose of forcing a positive prior (and equivalent) posterior mean in the context of non-identified correlation parameters.

APPENDIX L

Optimization Method Details

For the optimization method, the standard error is calculated as follows. Remember that the γ quantities of interest are functions of both the identified and non-identified model parameters. In the setting of random intercepts assuming conditional independence:

$\gamma_0 = f(\alpha_1, \beta_0, \beta_1, \sigma_{S1}, \sigma_{T0}, \sigma_{T1}, \rho_{11}, \rho_{10} = \rho_{11}\rho_T)$

$\gamma_1 = f(\alpha_1, \beta_0, \beta_1, \sigma_{S1}, \sigma_{T0}, \sigma_{T1}, \rho_{11}, \rho_{10} = \rho_{11}\rho_T)$

Let $\theta = (\alpha_1, \beta_0, \beta_1, \sigma_{S1}, \sigma_{T0}, \sigma_{T1}, \rho_{11})$ be the identified parameters.

Then $Var(\hat{\theta})$ is obtained from the Hessian.

Let $Var(\hat{\gamma}_.|\rho_{10}) = \frac{\partial \gamma_.}{\partial \theta}^T Var(\hat{\theta}) \frac{\partial \gamma_.}{\partial \theta}$

For J repeated draws of ρ_{10} within a single dataset, we calculate the final variance as

$Var(\hat{\gamma}_.) = \frac{1}{J}\Sigma_{j=1}^{J} Var(\hat{\gamma}_.|\rho_{10j}) + \frac{1}{J-1}\Sigma_{j=1}^{J}(\hat{\gamma}_.(\rho_{10j}) - \bar{\gamma}_.)^2$

where this idea applies for both γ_0 and γ_1. This method of drawing the nonidentified parameters from the prior many times and incorporating the uncertainty in the corresponding estimates is similar to a Rubin's Rules type formulation of combining within- and between-variability (Rubin, 1987).

APPENDIX M

Simulation Results for Sensitivity Analyses

13.1 Sensitivity Results of Sample Size

Setting Range of Data Generating Value Unif, No CI	γ_0 (-2.052, 1.466)	γ_0 SE	γ_0 SD	γ_1 (-0.182, 1.575)	γ_1 SE	γ_1 SD
n = 100, CI	0.169	0.450	0.242	0.474	0.225	0.119
n = 200, CI	0.040	0.425	0.202	0.534	0.213	0.097
n = 300, CI	0.002	0.408	0.155	0.552	0.205	0.078
n = 400, CI	-0.002	0.404	0.150	0.552	0.202	0.074
n = 500, CI	0.000	0.402	0.140	0.553	0.201	0.068
n = 1000, CI	0.005	0.388	0.096	0.550	0.194	0.047

Table 13.1: Simulation results of random intercept models comparing different sample sizes for the algorithm *MCMC, Obs Data, Random Effects, MH Beta Prior*. The true values are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices under conservative settings, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a Uniform(-1, 1) distribution with no conditional independence (CI) assumptions. Results shown for γ quantities are the posterior mean, average estimated standard error within simulation, and standard deviation of the point estimates across simulation replications. Obs Data represents the observed data algorithm (rather than an imputation scheme), and MH denotes when Metropolis Hastings steps were involved in the Markov Chain Monte Carlo (MCMC).

13.2 Sensitivity Analysis of Distributional Assumptions

Setting	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
Range of Data Generating Value Unif, No CI	(-2.052, 1.466)			(-0.182, 1.575)		
Correctly Specified	0.002	0.408	0.155	0.552	0.205	0.078
Misspecified	-0.201	0.472	0.210	0.722	0.235	0.106

Table 13.2: Simulation results of random intercept models for misspecified models (true random effects follow a skewed, joint Gamma distribution with rate and scale parameters 0.5) for the algorithm *MCMC, Obs Data, Random Effects, CI, MH Beta Prior*. The true values are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices under conservative settings, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a Uniform(-1, 1) distribution with no conditional independence (CI) assumptions. Results shown for γ quantities are the posterior mean, average estimated standard error within simulation, and standard deviation of the point estimates across simulation replications. Obs Data represents the algorithm that uses only the observed data (rather than an imputation scheme).

13.3 Sensitivity Analysis of Sample Size

Setting	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
Range of Data Generating Value Unif, No CI	(-1.508, 1.356)			(0.145, 1.577)		
n = 100, Time 1	0.194	0.450	0.354	0.704	0.215	0.162
n = 300, Time 1	0.079	0.266	0.250	0.777	0.128	0.124
n = 500, Time 1	0.059	0.207	0.204	0.786	0.100	0.091
n = 700, Time 1	0.010	0.174	0.147	0.804	0.083	0.062
n = 900, Time 1	-0.038	0.126	0.093	0.837	0.060	0.048

Table 13.3: Simulation results of random slopes models with varying sample sizes. We assess a delayed-start treatment design with a pre-treatment, baseline measurement of T, T_{BL} treated as a covariate and assume conditional independence. The true values are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a non-informative (Beta(3, 3)) distribution rescaled between -1 and 1.

13.4 Sensitivity Analysis of Random Slope Model Specification

Setting	γ_0	γ_0 SE	γ_0 SD	γ_1	γ_1 SE	γ_1 SD
Range of Data Generating Value Unif, No CI	(-1.508, 1.356)			(0.145, 1.577)		
Correctly Specified, Time 1	-0.038	0.126	0.093	0.837	0.060	0.048
Non-linear effect of time (Misspecified)	-0.098	0.133	0.153	0.860	0.063	0.074

Table 13.4: Simulation results of random slopes models for misspecified models where the true generating model includes a non-linear (quadratic) term for time that is not specified in the fitted model. We assess a delayed-start treatment design with a pre-treatment, baseline measurement of T, T_{BL} treated as a covariate and assume conditional independence. The true values are listed as the 2.5th and 97.5th quantiles of repeated draws from an infinite data setting of valid covariance matrices, meaning the identified parameters are set to their true generating values, and non-identified parameters are drawn from a non-informative (Beta(3, 3)) distribution rescaled between -1 and 1.

APPENDIX N

Defining Ideal CEP Curves for Time-to-Event Data

Model A Parameterization

First we consider models of the form

$$\lambda^0_{12}(t|\omega^0_{12i}) = \lambda^0_{12,0}(t)\exp(\kappa^0_{12}\omega^0_{12i}) \qquad \lambda^0_{13}(t|\omega^0_{13i}) = \lambda^0_{13,0}(t)\exp(\kappa^0_{13}\omega^0_{13i}) \tag{14.1}$$

$$\lambda^0_{23}(t|T_{12i}(0),\omega^0_{23i}) = \lambda^0_{23,0}(t-T_{12i}(0))\exp(\kappa^0_{23}\omega^0_{23i}+\theta^0_{23}T_{12i}(0))I(t>T_{12i}(0))$$

$$\lambda^1_{12}(t|\omega^1_{12i}) = \lambda^1_{12,0}(t)\exp(\kappa^1_{12}\omega^1_{12i}) \qquad \lambda^1_{13}(t|\omega^1_{13i}) = \lambda^1_{13,0}(t)\exp(\kappa^1_{13}\omega^1_{13i})$$

$$\lambda^1_{23}(t|T_{12i}(1),\omega^1_{23i}) = \lambda^1_{23,0}(t-T_{12i}(1))\exp(\kappa^1_{23}\omega^1_{23i}+\theta^1_{23}T_{12i}(1))I(t>T_{12i}(1))$$

and the eight scenarios described in the main text and shown in Table 14.1.

	$\lambda_{12}^0 = \lambda_{12}^1$	$\lambda_{13}^0 = \lambda_{13}^1$	$\lambda_{23}^0 = \lambda_{23}^1$	Surrogacy
Scenario 1	T	T	T	Null Case
Scenario 2	F	T	T	Perfect
Scenario 3	F	T	F	Partial
Scenario 4	F	F	T	Partial
Scenario 5	F	F	F	Partial
Scenario 6	T	F	F	Not a surrogate
Scenario 7	T	T	F	Not a surrogate
Scenario 8	T	F	T	Not a surrogate

Table 14.1: Eight possible scenarios of which pathways in the illness death models exhibit treatment effects based on the causal hazards. T denotes true and F denotes false. The right hand column represents an intuitive notion of whether S is a good surrogate for T.

The generating parameter values are shown with each plot. Each plot also shows the number of events observed in each arm for each transition: $n12(z), n13(z), n23(z)$. First we present the corresponding CEP plots for scenarios 1-8 when $\theta_{23}^1 = \theta_{23}^0 = 0$. We assume that $\omega_{13}^z = \omega_{23}^z$.

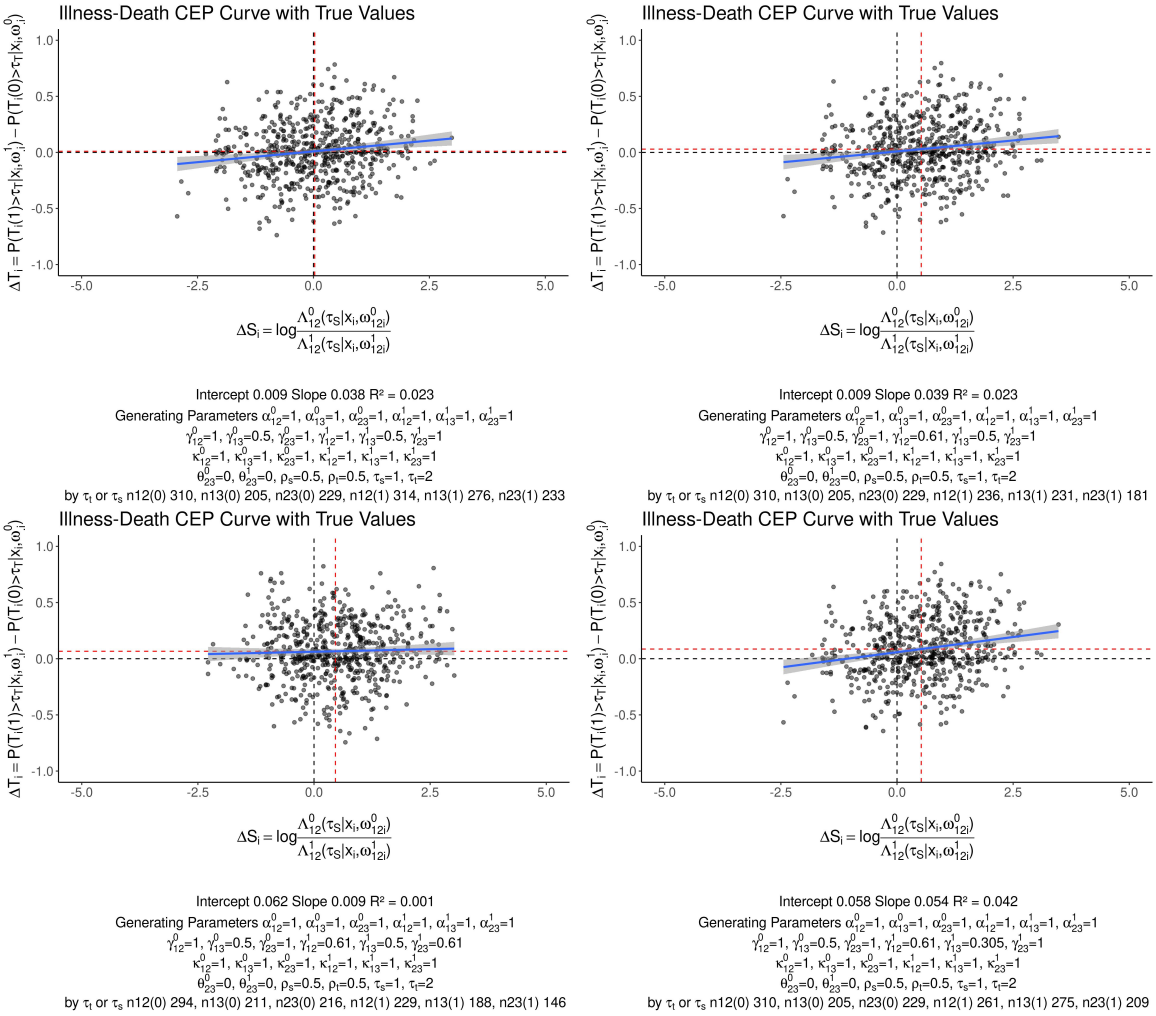


Figure 14.1: Scenarios 1-4: Null, perfect, and partial surrogates under the illness-death formulation.

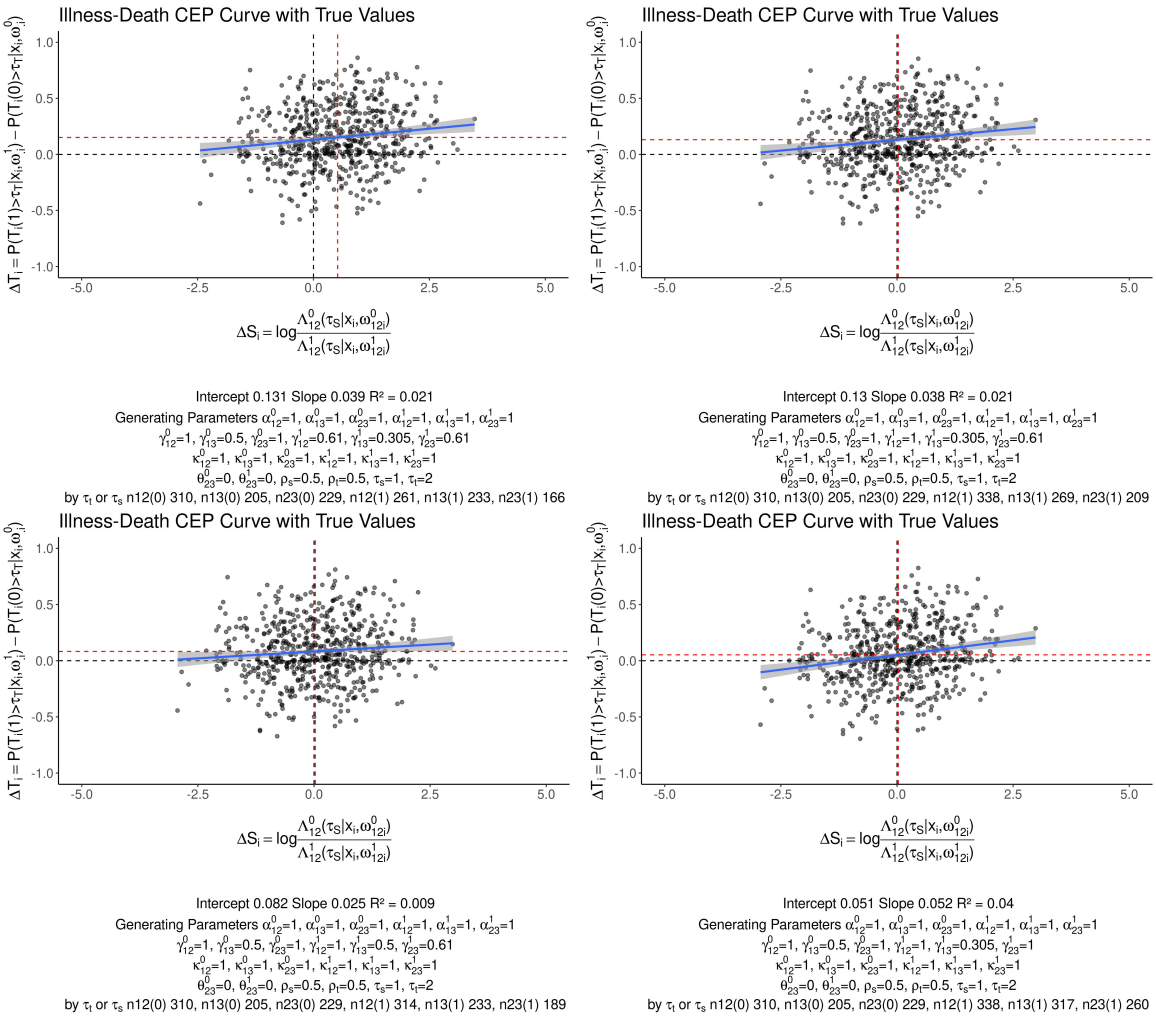


Figure 14.2: Scenarios 5-8: Partial and non-surrogates under the illness-death formulation.

We see that in the first two plots, when there are no treatment effects (Scenario 1) or only an effect through the surrogate endpoint (Scenario 2), the intercept is approximately 0 and the slope is positive. However, when there are treatment effects involved that do not go through the surrogate, the intercept becomes nonzero. In general, we note that the slope does not drastically change across the different scenarios above. Necessarily, the y-axis will always be constrained between -1 and 1 since it represents a difference in two probabilities. Further, this quantity on the y-axis is a relatively complex function of multiple model parameters. More exploration into what settings may induce a different slope is needed.

We find that changing the fixed values of ρ_S and ρ_T affects the spread of the points across

individuals. Consider this scenario comparing $\rho_S = \rho_T = 0$ to $\rho_S = \rho_T = 0.95$ in Figure 14.3.

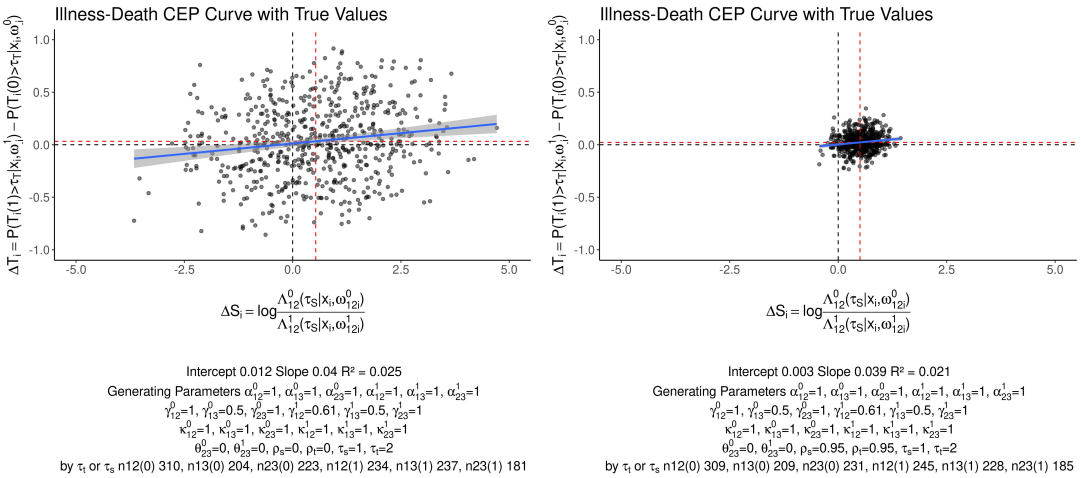


Figure 14.3: CEP curve when $\rho_S = \rho_T = 0$ versus when $\rho_S = \rho_T = 0.95$.

We consider when $\theta_{23}^z \neq 0$ so that the effect of time to the surrogate endpoint T_{12} affects time from S to T. It is apparent that T_{12}^z and T_{23}^z will be more highly correlated when $\theta_{23}^z < 0$, since longer times to S will be associated with a lower hazard (and longer time) to T afterward. It is our observation that the slope sometimes increases when $\theta_{23}^z < 0$, as shown in the figures below for Scenario 2. This behavior also depends on all of the other parameters, such as the magnitude of θ_{23}^z and value of τ_T.

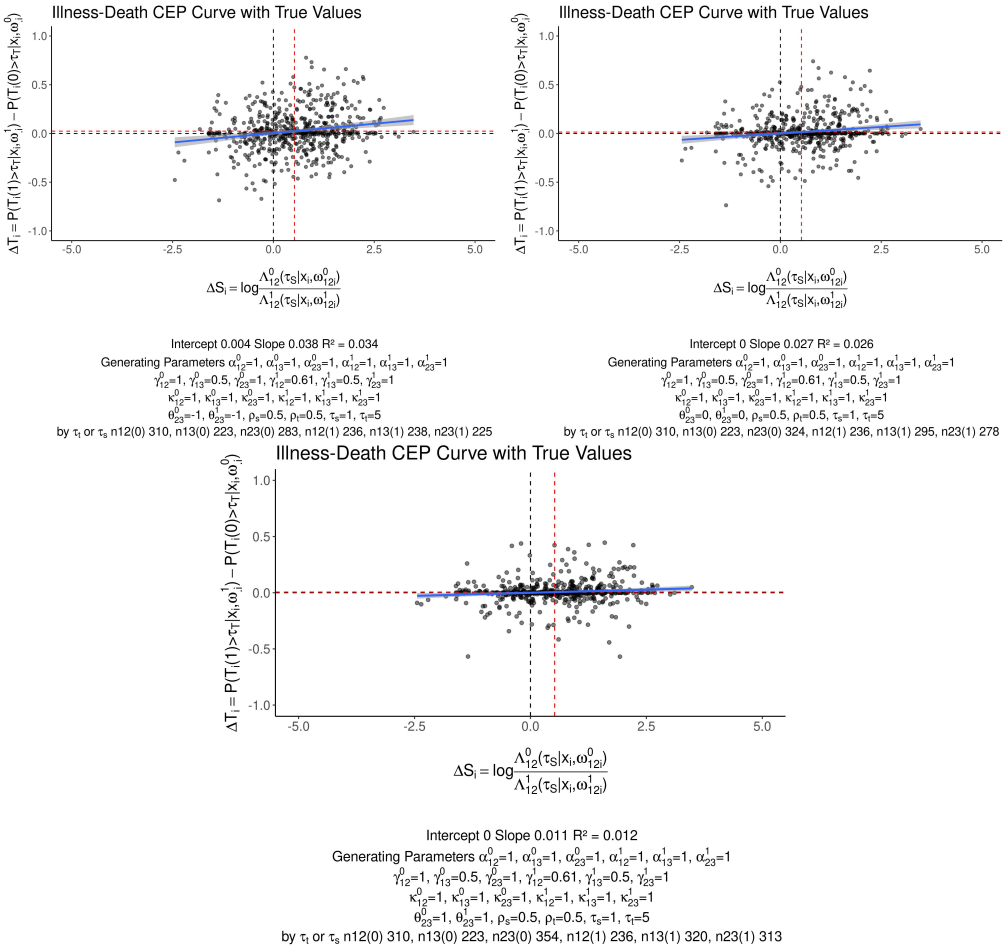


Figure 14.4: CEP curves comparing changing $\theta_{23}^z = -1, 0, 1$.

APPENDIX O

Prentice Approach Formulation and Relation to Proposed Illness-Death Method

Consider two models:

A) $\lambda(t)\exp(\phi_1 Z_i + \eta_1 X_i)$

B) $\lambda(t)\exp(\phi_2 Z_i + \eta_2 X_i + \omega I(t > S_i))$

Prentice approach: Compare ϕ_1 and ϕ_2 and evaluate whether $\phi_2 = 0$.

First expand B with interactions with Z_i.

$\lambda(t)\exp(\phi_2 Z_i + \eta_2^0 X_i I(Z_i = 0) + \eta_2^1 X_i I(Z_i = 1) + \omega^0(1 - Z_i)I(t > S_i) + \omega^1 Z_i I(t > S_i))$

for $Z = 0$, $= \lambda(t)\exp(\eta_2^0 X_i + \omega^0 I(t > S_i))$

for $Z = 1$, $= \lambda(t)\exp(\phi_2)\exp(\eta_2^1 X_i + \omega^1 I(t > S_i))$

Now call these λ_{13}^0, for $t < S_i$, $Z = 0$, $\lambda(t)\exp(\eta_2^0 X_i)$

λ_{23}^0, for $t \geq S_i$, $Z = 0$, $\lambda(t)\exp(\eta_2^0 X_i + \omega^0)$

λ_{13}^1, for $t < S_i$, $Z = 1$, $\lambda(t)\exp(\phi_2)\exp(\eta_2^1 X_i)$

λ_{23}^1, for $t \geq S_i$, $Z = 1$, $\lambda(t)\exp(\phi_2)\exp(\eta_2^1 + \omega^1)$

Generalize $\lambda_{13}^z, \lambda(t) \to \lambda_{13}^0(t), \lambda(t)\exp(\phi_2) \to \lambda_{13}^1(t)$ and add frailties $\omega_{13i}^0, \omega_{13i}^1$

$\Rightarrow \lambda_{13}^z(t)\exp(\eta_2^Z X_i + \omega_{13i}^z)$

Generalize $\lambda_{23}^z, \lambda(t)\exp(\omega^0) \to \lambda_{23}^0(t), \lambda(t)\exp(\omega^1 + \phi_2) \to \lambda_{23}^1(t)$ and add frailties $\omega_{23i}^0, \omega_{23i}^1$

$\Rightarrow \lambda_{23}^z(t)\exp(\eta_2^Z X_i + \omega_{23i}^z)I(t > S_i^z)$

We can change the scale to $\tau = t - S_i^z$ and add dependence on S

$\lambda_{23}^z(\tau)\exp(\omega_{23i}^z + \eta_2^z X_i + \theta^z S_i^z)$

Model A is some combination of $\lambda_{12}, \lambda_{13}, \lambda_{23}$. We replace A by a specification of $\lambda^z_{12}(t)$

$\Rightarrow \lambda^z_{12}(t) \exp(\eta^z_1 X_i + \omega^z_{12i})$

We can restrict either $\omega^z_{23i} = \kappa\omega^z_{13i}$ and $\theta^z \neq 0$ or $\omega^z_{23i} = \kappa_1\omega^z_{12i} + \kappa_2\omega^z_{13i}$ and $\theta^z = 0$ and see the resemblance to our models proposed in the text.

APPENDIX P

Likelihood Contributions for Illness-Death Model Parameters

In this appendix, we provide the likelihood contributions for the parameters in Model A for use in the Bayesian estimation strategy described in Chapter IV Section 4. For ease of notation, we will consider other covariates later.

Contribution for regression coefficients that we consider here, θ_{23}, κ_{jk}

Let $\beta_{12i} = \exp(\kappa_{12} w_{12i})$. From the likelihood, the contribution for the κ_{12} parameter comes from $\Pi_i^{n_z} \exp(-\Lambda_{12}(T_{12i}))\lambda_{12}(T_{12i})^{\delta_{Si}}$

$\propto \Pi_i^{n_z} \exp(-\beta_{12i}\Lambda_{120}(T_{12i})) \exp(\log(\beta_{12i}) \times \delta_{Si})$

$\propto \exp(-\gamma_{12}\Sigma_i^{n_z}\beta_{12i}T_{12i}^{\alpha_{12}} + \Sigma_i^{n_z}\delta_{Si}\kappa_{12}\omega_{12i})$

Now consider the likelihood for parameters in λ_{13}, κ_{13}. Let n_{13} be the number of individuals who do not experience S, n_{23} the number who do experience S, and $\beta_{13i} = \exp(\kappa_{13} w_{13i})$.

From the likelihood, $\Pi_i^{n_{13}} \exp(-\Lambda_{13}(T_{13i}))\lambda_{13}(T_{13i})^{\delta_{Ti}} \times \Pi_i^{n_{23}} \exp(-\Lambda_{13}(T_{12i}))$

$\propto \Pi_i^{n_{13}} \exp(-\beta_{13i}\Lambda_{130}(T_{13i})) \exp(\log(\beta_{13i}) \times \delta_{Ti}) \times \Pi_i^{n_{23}} \exp(-\Lambda_{13}(T_{12i}))$

$\propto \exp(-\Sigma_i^{n_{13}}(\beta_{13i}\ \gamma_{13}T_{13i}^{\alpha_{13}} + \kappa_{13}\omega_{13i}\delta_{Ti})) \exp(-\Sigma_i^{n_{23}}\gamma_{13}T_{12i}^{\alpha_{13}}\beta_{13i})$

Now consider the likelihood for parameters in λ_{23}, θ_{23} and κ_{23}. Let $\beta_{23i} = \exp(\kappa_{23} w_{23i} + \theta_{23} T_{12i})$.

From the likelihood, $\Pi_i^{n_{23}} \exp(-\Lambda_{23}(T_{23i}))\lambda_{23}(T_{23i})^{\delta_{Ti}}$

$\propto \Pi_i^{n_{23}} \exp(-\beta_{23i}\Lambda_{230}(T_{23i})) \exp(\log(\beta_{23i}) \times \delta_{Ti})$

$\propto \exp(-\gamma_{23}\Sigma_i^{n_{23}}\beta_{23i}T_{23i}^{\alpha_{23}} + \Sigma_i^{n_{23}}\delta_{Ti}(\kappa_{23}\omega_{23i} + \theta_{23}T_{12i}))$

The contributions for the frailty terms ω_{jki} are similar, except they do not have the sum or product over all individuals. The contributions for shape and scale parameters α_{jk} and γ_{jk} are also similar, specifically relying on the respective quantities

$$\propto \exp(-\gamma_{12}\Sigma_i^{n_z}\beta_{12i}T_{12i}^{\alpha_{12}})$$

$$\propto \exp(-\gamma_{13}\Sigma_i^{n_{13}}\beta_{13i}T_{13i}^{\alpha_{13}})\exp(-\gamma_{13}\Sigma_i^{n_{23}}T_{12i}^{\alpha_{13}}\beta_{13i})$$

$$\propto \exp(-\gamma_{23}\Sigma_i^{n_{23}}\beta_{23i}T_{23i}^{\alpha_{23}})$$

for $jk = 12, 13,$ and 23. Since T_{23} is not observed for all individuals, only individuals who have $\delta_{Si} = 1$ (so that the time T_{23i} exists) contribute to the last likelihood component. This sum is calculated over these n_{23} individuals for a given treatment arm. A closed-form posterior distribution for a general scale parameter γ_{jk} (not considering the illness-death framework) is given in Sahu et al. (1997) that may be applied under a Gamma prior with a Weibull baseline hazard.

APPENDIX Q

Simulation Results for Illness-Death Model Parameters

Here we provide a subset of results for illness-death proposed model assuming that the scale parameters are fixed. In these simulations, θ_{23} is fixed to its true value to simplify the model, $\rho_s = \rho_T = 0.5$, and $\tau_S = 0.5$ and $\tau_T = 1.5$.

	γ_0	γ_1
True Value*	0	0.035
Scenario 1: Estimates	0.012	0.021
SE	0.035	0.020
SD	0.049	0.023
True Value*	0	0.040
Scenario 2: Estimates	0.021	0.028
SE	0.032	0.018
SD	0.041	0.021
True Value*	-0.089	0.054
Scenario 3: Estimates	-0.046	0.028
SE	0.034	0.018
SD	0.043	0.022
True Value*	-0.081	0.057
Scenario 7: Estimates	-0.066	0.021
SE	0.037	0.019
SD	0.052	0.023

Table 17.1: Simulation results from illness-death models and estimated validation quantities. This table shows the posterior mean, average estimated standard error (SE), and the standard deviation (SD) of the posterior means across simulation replications.
In these calculations, the scale parameters are fixed.
*Based on empirical calculations from a larger sample size over many replications

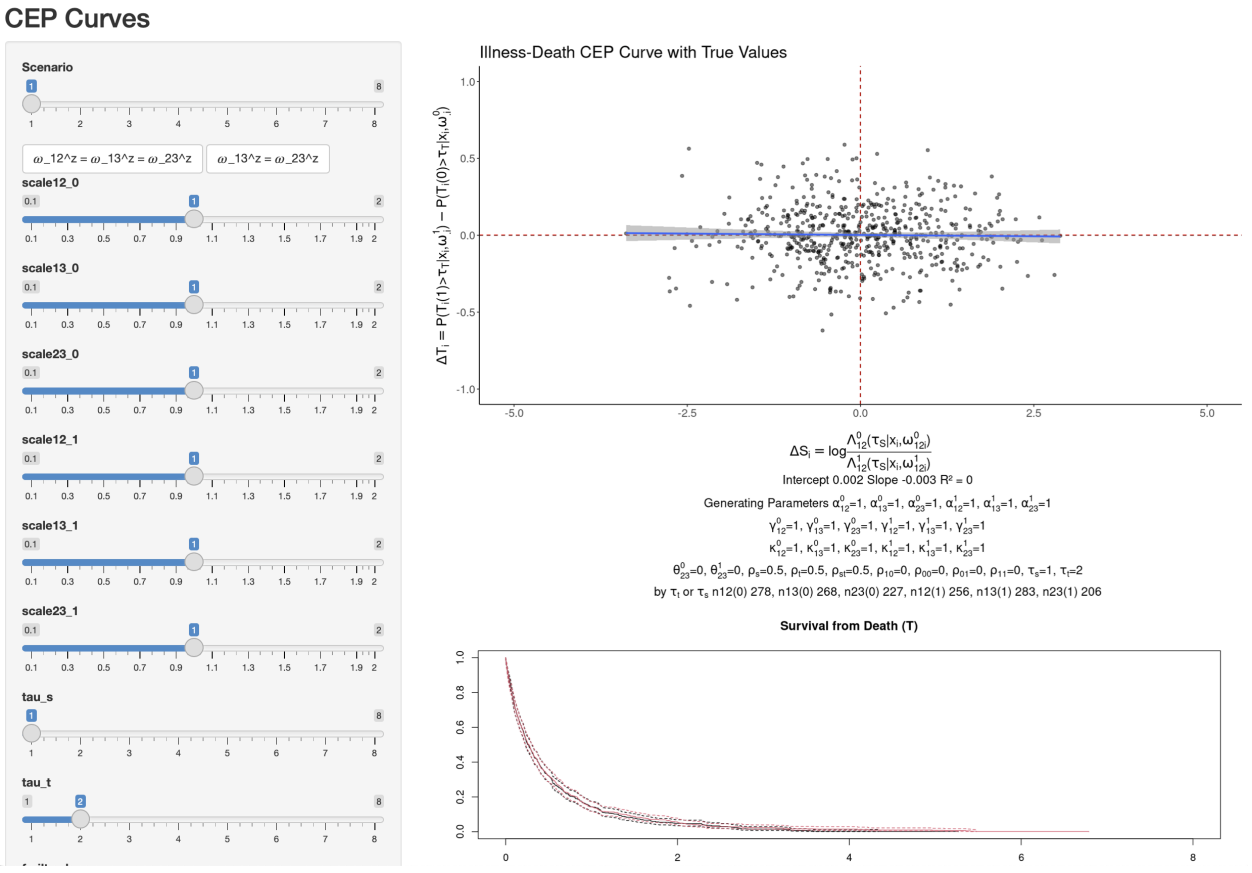
	κ_{12}^0	κ_{13}^0	κ_{23}^0	κ_{12}^1	κ_{13}^1	κ_{23}^1
True Value	1	1	1	1	1	1
Scenario 1: Estimates	1.00	1.00	0.96	1.00	1.00	1.04
SE	-	-	0.14	-	-	0.16
SD	-	-	0.15	-	-	0.15
Scenario 2: Estimates	1.00	1.00	0.97	1.00	1.00	1.07
SE	-	-	0.16	-	-	0.16
SD	-	-	0.15	-	-	0.15
Scenario 3: Estimates	1.00	1.00	0.97	1.00	1.00	1.03
SE	-	-	0.16	-	-	0.16
SD	-	-	0.17	-	-	0.18
Scenario 7: Estimates	1.00	1.00	0.95	1.00	1.00	1.04
SE	-	-	0.16	-	-	0.16
SD	-	-	0.16	-	-	0.15

Table 17.2: Simulation results from illness-death models of regression coefficients when scale parameters are fixed.

APPENDIX R

Rshiny App for Illness-Death Model Parameters

Here we provide more details about the Rshiny app available at https://emilyroberts.shinyapps.io/id_cep_parameters/ including a snapshot of the user-facing interface at this link.



On the left hand panel, users can input several parameters to investigate, including scenarios 1-8, scale parameters of the Weibull distribution, τ_T and τ_S at which times the surrogate is evalu-

ated, and several assumptions about the frailty terms. These include which frailties are equal (via clickable buttons) and the values of several ρ parameters (via sliding bars) that control how the frailties are correlated. The frailties are based on the parameterization

$$\begin{pmatrix} \omega_{12i}^0 \\ \omega_{12i}^1 \\ \omega_{13i}^0 \\ \omega_{13i}^1 \\ \omega_{23i}^0 \\ \omega_{23i}^1 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_S & \rho_{00} & \rho_{01} & \rho_{S1} & \rho_{S2} \\ & 1 & \rho_{10} & \rho_{11} & \rho_{S3} & \rho_{S4} \\ & & 1 & \rho_T & \rho_{T1} & \rho_{T2} \\ & & & 1 & \rho_{T3} & \rho_{T4} \\ & & & & 1 & \rho_{ST} \\ & & & & & 1 \end{pmatrix} \right)$$

The right hand side of the app provides the calculated CEP curve and the Kaplan-Meier plots for time to T (through either pathway of the illness death model) and time to S (where S is censored if T occurs first). In the top CEP curve, the red dashed line indicates the average ΔS and ΔT. The numbers under the figure indicate the parameter values used in the plot and the number of individuals who experienced $T_{12}(z)$ by τ_S or $T_{23}(z)$ and $T_{13}(z)$ by τ_T as $n12(z), n23(z), n13(z)$, respectively. With the Kaplan-Meier plots, we provide the difference in the heights of the curves at τ_S and τ_T.

BIBLIOGRAPHY

Aalen, O. O., Cook, R. J., and Røysland, K. (2015). Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21(4):579–593.

Agniel, D. and Parast, L. (2020). Evaluation of longitudinal surrogate markers. *Biometrics*, 77(2):477–489.

Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, 18(13):1707–1732.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. G., and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(8):931–945.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. G., and Vangeneugden, T. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics*, 60(4):845–53.

Alonso, A., Van der Elst, W., and Meyvisch, P. (2017). Assessing a surrogate predictive value: a causal inference approach. *Statistics in Medicine*, 36(7):1083–1098.

Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958.

Balan, T. A. and Putter, H. (2019). frailtyEM: An R package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90(7):1–29.

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage with application to shrinkage. *Statistica Sinica*, pages 1281–1311.

Beesley, L. J. and Taylor, J. M. (2019). EM algorithms for fitting multistate cure models. *Biostatistics*, 20(3):416–432.

Beyer, U., Dejardin, D., Meller, M., Rufibach, K., and Burger, H. U. (2020). A multistate model for early decision-making in oncology. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 62(3):550–567.

Bhadra, A., Rao, A., and Baladandayuthapani, V. (2018). Inferring network structure in non-normal and mixed discrete-continuous genomic data. *Biometrics*, 74(1):185–195.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6(1):1–9.

Burzykowski, T. and Buyse, M. (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 5(3):173–186.

Clayton, D. G. (1991). A monte carlo method for bayesian inference in frailty models. *Biometrics*, pages 467–485.

Comment, L., Mealli, F., Haneuse, S., and Zigler, C. (2019). Survivor average causal effects for continuous time: a principal stratification approach to causal inference with semicompeting risks. *arXiv preprint*, page 1902.09304.

Conlon, A., Taylor, J., and Elliott, M. (2017a). Surrogacy assessment using principal stratification and a gaussian copula model. *Statistical methods in medical research*, 26(1):88–107.

Conlon, A. S., Taylor, J. M. G., and Elliott, M. R. (2014a). Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics*, 15(2):266–283.

Conlon, A. S. C., Taylor, J. M. G., Li, Y., Diaz-Ordaz, K., and Elliott, M. R. (2017b). Links between causal effects and causal association for surrogacy evaluation in a gaussian setting. *Statistics in Medicine*, 36(7):4243–4265.

Conlon, A. S. C., Taylor, J. M. G., and Sargent, D. J. (2014b). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine*, 33(10):1750–1766.

Daniels, M. J., Roy, J. A., Kim, C., Hogan, J. W., and Perri, M. G. (2012). Bayesian inference for the causal effect of mediation. *Biometrics*, 68(4):1028–1036.

de Castro, M., Chen, M. H., and Zhang, Y. (2015). Bayesian path specific frailty models for multi-state survival data with applications. *Biometrics*, 71(3):760–771.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671.

Elliott, M. R., Conlon, A. S., Li, Y., Kaciroti, N., and Taylor, J. M. G. (2015). Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics*, 16(2):400–412.

Emura, T., Nakatochi, M., Murotani, K., and Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta- analysis. *Statistical Methods in Medical Research*, 26:2649–2666.

Fleischer, F., Gaschler-Markefski, B., and Bluhmki, E. (2009). A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine*, 28(21):2669–2686.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62(4):1161–1169.

Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11(2):167–78.

Gabriel, E. E. and Follmann, D. (2016). Augmented trial designs for evaluation of principal surrogates. *Biostatistics*, 17(3):453–467.

Gabriel, E. E. and Gilbert, P. B. (2014). Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics*, 15(2):251–265.

Gabriel, E. E., Sachs, M. C., Daniels, M. J., and Halloran, M. E. (2019). Optimizing and evaluating biomarker combinations as trial-level general surrogates. *Statistics in Medicine*, 38(7):1135–1146.

Gabriel, E. E., Sachs, M. C., and Gilbert, P. B. (2015). Comparing and combining biomarkers as principal surrogates for time-to-event clinical endpoints. *Statistics in Medicine*, 34(3):76–105.

Gao, X. (2012). *Causal Modeling with Principal Stratification to Assess Effects of Treatment with Partial Compliance, Noncompliance, and Principal Surrogacy in Longitudinal and Time-to-Event Settings*. PhD thesis, University of Michigan.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient metropolis jumping rules. *Bayesian Statistics*, 5:42.

Ghosh, D. (2009). On assessing surrogacy in a single trial setting using a semicompeting risks paradigm. *Biometrics*, 65(2):521–529.

Ghosh, D. (2012a). A causal framework for surrogate endpoints with semi-competing risks data. *Statistics & Probability Letters*, 82(11):1898–1902.

Ghosh, D. (2012b). A causal framework for surrogate endpoints with semi-competing risks data. *Statistics & probability letters*, 82(11):1898–1902.

Ghosh, D., Taylor, J. M., and Sargent, D. J. (2012). Meta-analysis for surrogacy: Accelerated failure time models and semicompeting risks modeling. *Biometrics*, 68(1):226–232.

Gilbert, P. and Hudgens, M. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154.

Gilbert, P. B. and Huang, Y. (2016). Predicting overall vaccine efficacy in a new setting by re-calibrating baseline covariate and intermediate response endpoint effect modifiers of type-specific vaccine efficacy. *Epidemiologic Methods*, 5(1):93–112.

Gilbert, P. B., Qin, L., and Self, S. G. (2008). Evaluating a surrogate endpoint at three levels with application to vaccine development. *Statistics in Medicine*, 62(4):4758–4778.

Gran, J. M., Lie, S. A., Øyeflaten, I., Borgan, Ø., and Aalen, O. (2015). Causal inference in multi-state models sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, 15(1):1–16.

Gustafson, P. (2009). What are the limits of posterior distributions arising from nonidentified models and why should we care? *Journal of the American Statistical Association*, 104(488):1682–1695.

Halloran, M. E., Préziosi, M. P., and Chu, H. (2003). Estimating vaccine efficacy from secondary attack rates. *Journal of the American Statistical Association*, 98(461):38–46.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13.

Hernán, M. A. and Robins, J. M. (2010). *Causal inference*. Boca Raton, FL:: CRC.

Hsu, J. Y., Kennedy, E. H., Roy, J. A., Stephens-Shields, A. J., Small, D. S., and Joffe, M. M. (2015). Surrogate markers for time-varying treatments and outcomes. *Clinical Trials*, 12(4):309–316.

Huang, Y., Gilbert, P. B., and Wolfson, J. (2013). Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics*, 69(2):301–309.

Hudgens, M. G. and Halloran, M. E. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association*, 101(473):51–64.

Huuang, Y. T. (2021). Causal mediation of semicompeting risks. *Biometrics*, 77(4):1143–1154.

Jackson, W. C., Tang, M., Schipper, M., Sandler, H. M., Zumsteg, Z. S., Efstathiou, J. A., Shipley, W. U., Seiferheld, W., Lukka, H., Bahary, J. P., Zietman, A. L., Pisansky, T. M., Zeitzer, K. L., Hall, W. A., Dess, R. T., Lovett, R. D., Balogh, A., Feng, F. Y., and Spratt, D. E. (2020). Metastasis-free survival, but not biochemical failure, is a strong surrogate endpoint for overall survival in recurrent prostate cancer: Analysis of NRG oncology/RTOG 9601. *International Journal of Radiation Oncology, Biology, Physics*, 108(3):S63–S64.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538.

Kemp, R. and Prasad, V. (2017). Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC medicine*, 15(1):1–7.

Kim, C., Daniels, M. J., Marcus, B. H., and Roy, J. A. (2017). A framework for bayesian nonparametric inference for causal effects of mediation. *Biometrics*, 73(2):401–409.

Li, Y. and Taylor, J. M. G. (2010). Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine*, 29(18):1875–1889.

Li, Y. and Zhang, Q. (2015). A weibull multi-state model for the dependence of progression-free survival and overall survival. *Statistics in Medicine*, 34(17):2497–2513.

Little, R. and Rubin, D. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21(1):121–145.

Luedtke, A. and Wu, J. (2020). Efficient principally stratified treatment effect estimation in crossover studies with absorbent binary endpoints. *Journal de la Société Française de Statistique*, 161(1):176–200.

Ma, Y., Roy, J., and Marcus, B. (2011). Causal models for randomized trials with two active treatments and continuous compliance. *Statistics in Medicine*, 30(19):2349–2362.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26(4):833–855.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549.

Maurer, W. and Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research*, 5(4):311–320.

Meller, M., Beyersmann, J., and Rufibach, K. (2019). Joint modeling of progression-free and overall survival and computation of correlation measures. *Statistics in Medicine*, 38(22):4270–4289.

Mendell, J. R., Sahenk, Z., Lehman, K., Nease, C., Lowes, L. P., Miller, N. F., et al. (2020). Assessment of systemic delivery of raavrh74. mhck7. micro-dystrophin in children with duchenne muscular dystrophy: A nonrandomized controlled trial. *JAMA Neurology*, 77(9):1122–1131.

Muntoni, F., Domingos, J., Manzur, A. Y., Mayhew, A., Guglieri, M., U. K. NorthStar Network, and Ward, S. J. (2019). Categorising trajectories and individual item changes of the north star ambulatory assessment in patients with duchenne muscular dystrophy. *PloS One*, 14:9.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Omori, Y. and Johnson, R. A. (1993). The influence of random effects on the unconditional hazard rate and survival functions. *Biometrika*, 80(4):910–914.

O'Quigley, J. and Flandre, P. (2012). Discussion on "meta-analysis for surrogacy: Accelerated failure time models and semicompeting risks modeling". *Biometrics*, 68(1):242–245.

Parast, L., Cai, T., and Tian, L. (2016a). Nonparametric estimation of the proportion of treatment effect explained by a surrogate marker using censored data. *Technical Report*.

Parast, L., Cai, T., and Tian, L. (2017). Evaluating surrogate marker information using censored data. *Statistics in Medicine*, 36(11):1767–1782.

Parast, L., McDermott, M. M., and Tian, L. (2016b). Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in Medicine*, 35(10):1637–1653.

Parmar, M. K., Barthel, F. M. S., Sydes, M., Langley, R., Kaplan, R., Eisenhauer, E., and Qian, W. (2008). Speeding up the evaluation of new agents in cancer. *Journal of the National Cancer Institute*, 100(17):1204–1214.

Pearl, J. and Bareinboim, E. (2011). Transportability across studies: A formal approach. Technical report, California University of Las Angeles Department of Computer Science.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554.

Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4):431–40.

Préziosi, M. P. and Halloran, M. E. (2003). Effects of pertussis vaccination on disease: vaccine efficacy in reducing clinical severity. *Clinical Infectious Diseases*, 37(6):772–779.

Price, B. L., Gilbert, P. B., and van der Laan, M. J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4):1271–1281.

Putter, H. (2011). *Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package*. Leiden University Medical Center, Department of Medical Statistics and Bioinformatics. Online Tutorial, Leiden.

Putter, H. and van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical Methods in Medical Research*, 24(6):675–692.

Qin, L., Gilbert, P. B., Follmann, D., and Li, D. (2008). Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the cox model. *The Annals of Applied Statistics*, 2(1):386.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T., and Bijnens, L. (2003). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics*, 30(2):235–247.

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. (2014). Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 29(4):596.

Robert, C. P. and Casella, G. (2004). The metropolis—hastings algorithm. In *Monte Carlo statistical methods*, pages 267–320. Springer.

Roberts, E. K., Elliott, M. R., and Taylor, J. M. (2021). Incorporating baseline covariates to validate surrogate endpoints with a constant biomarker under control arm. *Statistics in Medicine*, 40(29):6605–6618.

Rondeau, V. and Gonzalez, J. R. (2005). Frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer methods and programs in biomedicine*, 80(2):154–164.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Sahu, S. K., Dey, D. K., Aslanidou, H., and Sinha, D. (1997). A weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, 3(2):123–137.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization conditional on covariates, with application to HIV vaccine trials. *Biometrics*, 62(2):332–342.

Shipley, W. U., Seiferheld, W., Lukka, H. R., et al. (2017). Radiation with or without antiandrogen therapy in recurrent prostate cancer. *New England Journal of Medicine*, 376(5):417–428.

Sofeu, C. L., Emura, T., and Rondeau, V. (2019). One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints. *Statistics in Medicine*, 38(16):2928–2942.

Sofeu, C. L., Emura, T., and Rondeau, V. (2020). A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 63(2):423–446.

Song, P. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320.

Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S., and Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 59(6):1277–1300.

Sydes, M. R., Parmar, M. K., James, N. D., Clarke, N. W., Dearnaley, D. P., Mason, M. D., and Royston, P. (2009). Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the mrc stampede trial. *Trials*, 10(1):1–16.

Tanaka, S., Matsuyama, Y., and Ohashi, Y. (2017). Validation of surrogate endpoints in cancer clinical trials via principal stratification with an application to a prostate cancer trial. *Statistics in Medicine*, 36(19):2963–2977.

Taylor, J. M. G., Conlon, A. S., and Elliott, M. R. (2015). Surrogacy assessment using principal stratification with multivariate normal and gaussian copula models. *Clinical Trials*, 12(4):317–322.

Vandenberghe, S., Duchateau, L., Slaets, L., Bogaerts, J., and Vansteelandt, S. (2018). Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Statistical Methods in Medical Research*, 27(11):3367–3385.

VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics*, 69(3):561–565.

Weir, I. R., Rider, J. R., and Trinquart, L. (2021). Counterfactual mediation analysis in the multistate model framework for surrogate and clinical time-to-event outcomes in randomized controlled trials. *Pharmaceutical Statistics*, 21(1):163–175.

Wen, S., Huang, X., Frankowski, R. F., Cormier, J. N., and Pisters, P. (2016). A bayesian multivariate joint frailty model for disease recurrences and survival. *Statistics in Medicine*, 35(26):4794–4812.

Wolfson, J. and Gilbert, P. (2010). Statistical identifiability and the surrogate endpoint problem with application to vaccine trials. *Biometrics*, 66(4):1153–1161.

Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830.

Xie, W., Regan, M. M., Buyse, M., et al. (2017). Metastasis-free survival is a strong surrogate of overall survival in localized prostate cancer. *Journal of Clinical Oncology*, 35(27):3097–3104.

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics*, 66(3):716–725.

Xu, Y., Scharfstein, D., Moeller, P., and Daniels, M. (2020). A bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics*, 23(1):34–49.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, Y., Chen, M. H., Ibrahim, J. G., Zeng, D., Chen, Q., Pan, Z., and Xue, X. (2014). Bayesian gamma frailty models for survival data with semi-competing risks and treatment switching. *Lifetime Data Analysis*, 20(1):76–105.

Zhuang, Y., Huang, Y., and Gilbert, P. B. (2019). Simultaneous inference of treatment effect modification by intermediate response endpoint principal strata with application to vaccine trials. *The International Journal of Biostatistics*, 16(1):1.

Zigler, C. M. and Belin, T. R. (2012). A bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics*, 68(3):922–932.