Methods for Improving Inference in Clinical Outcomes

By

Duncan C. Rotich

B.Sc. Applied Statistics, Moi University, 2011

M.S. Applied Mathematics, University of Central Missouri 2015

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. Jeffrey Thompson, Co-Chair

Dr. Jonathan Mahnken, Co-Chair

Dr. Jo Wick

Dr. Byron Gajewski

Dr. Won Choi

Date defended: April 27, 2020

The Dissertation Committee for Duncan C. Rotich certifies
that this is the approved version of the following dissertation:

Methods for Improving Inference in Clinical Outcomes

Dr. Jeffrey Thompson, Co-Chair

Dr. Jonathan Mahnken, Co-Chair

Dr. Jo Wick, Program Director

Date approved: May 08, 2020

Abstract

Advances in technology have allowed for the collection of diverse data types along with evolution in computer algorithms. This dissertation focuses on the development and application of novel methodologies to model and improve inference on clinical outcomes. First, a new prognostic approach of modeling time-to-event data using Bayesian Networks (BNs) is developed and illustrated using publicly available cancer data. This approach allows for flexible modeling of different structural relationships that might exist between variables at different periods, hence, improving our understanding of critical prognostic factors that can inform patient care and development of targeted interventions. As a prognostic model, BNs demonstrated better or comparable performance as compared to other equivalent models for bladder and lung cancer data. In this dissertation, we also reviewed application of predictive modeling algorithms in randomized clinical trials (RCTs). RCTs are costly and time-consuming. Predictive modeling has the potential to mitigate challenges associated with clinical trial failures and facilitate efficient clinical trial conduct in areas such as patient recruitment, trial optimization, and safety & efficacy evaluations. Finally, we present a new approach for estimating causal treatment effect in RCTs that are prone to post-randomization intercurrent events (ICE). Examples of ICE include treatment switch, treatment discontinuation, or adverse events. Here, we adopt the principal stratification framework where we first predict the latent strata membership using baseline covariates and then estimated causal treatment effects using appropriate stratum having a homogeneous group of subjects. Using simulations, our approach demonstrated a better performance in estimating treatment effects as compared to the standard intent-to-treat (ITT) strategy.

Acknowledgements

I owe a debt of gratitude to a lot of people that have become a significant influence during my academic career development. I would not have made it this far without their support and encouragement.

First of all, I would like to express my deepest gratitude to my chairs for their guidance and unwavering support during my graduate research. Dr. Jeffrey Thompson for always being there for me and developing me into who I am today through his advising, mentorship, and support. For challenging and guiding me on the right path during my regular meetings. I have learned a lot from you and I feel so much indebted for your academic, professional, and personal mentorship. Dr. Jonathan Mahnken for guiding me during the early phase of my research, always caring about my progress, the project collaborations as well as recommending me to Dr. Thompson for dissertation project lead.

I am extremely grateful to my other dissertation committee members, first, for willing to be part of this journey: Dr. Jo Wick, Dr. Byron Gajewski, and Dr. Won Choi. I feel blessed to have all of you as my committee and I sincerely appreciate all the support and recommendation appertaining to my research. I can't thank you all enough.

Most of all, my deepest gratitude to Dr. Matthew Mayo (Department Chair) for believing in me to pursue this program. Financial support and creating a conducive learning environment in the department. Your leadership in the department is a great inspiration and I am lucky to have gone through this program under your leadership.

Special thanks to Dr. Jo Wick for all the advising throughout the course of this program, the encouragements and always wishing the best out of us. All the recommendations on professional development workshops and conferences.

Also, thanks to all my previous advisors and supervisors for providing me with a great

platform to develop my research skills from Dr. John Keighley during my first year in the graduate program, Dr. Jianghua He (Wendy), Shana Palla, Alvin Beltramo. I have picked up a lot of skills from each one of you. Thanks to all the faculty in the department for instilling the knowledge in me. All the staff, for their selfless behind the scene support, Rosemary Morrow, Mandy Rametta, Lori McElgunn, and Megan Tremblay.

Many thanks to the OMICs working group led by Dr. Devin Koestler and Dr. Jeffrey Thompson. The collaborative environment created by this group has been invaluable and all the weekly meetings, presentations, discussion and their constructive feedback accelerated my research and I am so grateful to this team for the influence they had on my research. (OMICs members: Richard Meier PhD, Lisa Neums, Dong Pei PhD, Qing Xia, Shachi Patel, Rosalyn Henn, Shelby Bell, Bo Zhang, Whitney Shae, Dr. Prabhakar Chalise, Emily Nissen, Dr. Lynn Chollet Hinton, Dr. Jinxiang Hu, Nanda Kumar Yellapu PhD, and Sefan Graw PhD (graduated member)).

Thanks to Alex Karanevich PhD (former student and a great friend) for your friendship and the idea on a collaborative project, "Predictive Modeling in Clinical Trials: A Narrative Review", which is part of this dissertation. My colleagues Lisa Neums, Shachi Patel, Richard Meier PhD, Stefan Graw PhD, Jiawei Duan PhD (Allan), Guangyi Gao (GiGi), Lauren Clark, Chuanwu Zhang, Pengcheng Lu PhD and many friends who have had a significant impact during this journey. I will forever be grateful to all of you to whom I have met at different stages of my educational and professional development.

I am also grateful to my family here in the US and in Kenya for their love, care, encouragement and support. Sharon Katam, Elizabeth Lagat, Natalie Halpin, Lem Shattuck and my siblings in Kenya.

Lastly, I would like to acknowledge the support from these sources that funded most

part of my research: The National Institutes of Health grant P20GM130423; the National Cancer Institute Cancer Center Support Grant P30 CA168524; the Kansas IDeA Network of Biomedical Research Excellence (KINBRE) Bioinformatics Core; and supported in part by the National Institute of General Medical Science award P20GM103418.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diagnosis and prognosis are fundamental aspects of medical practice that are used to inform a lot of decisions from treatment regimens to the investigation of new effective interventions. Most often, informed decisions are made on the likelihood of disease presence (diagnosis) or predicting future risk of occurrence of a specific event (prognosis) (Collins et al., 2015; Moons et al., 2015) based on observed patient characteristics. Lately, there has been an increased interest in developing optimal patient-centered therapeutic approaches tailored to meet individual patient's needs termed as personalized, stratified, or precision medicine (Erikainen & Chan, 2019; Juengst & McGowan, 2018; Personalized Medicine Coalition, 2014; Abrahams, 2008). These have been driven by advances in data collection technologies on high-dimensional data as well as evolution in computing algorithms. However, there still exists a need for more robust methods capable of handling diverse data types collected from patients in order capture patient heterogeneity to improve inference on clinical outcomes and promote development of novel therapeutic interventions in order to achieve precision medicine.

Analysis of time-to-event (TTE) data is the center stage of most biomedical cancer studies. The most common and widely used modeling approach for time-to-event data involves the Cox proportional hazards model (Cox, 1972). This model assumes a constant multiplicative effect of covariates on the hazards function over time. However, violation of this assumption might lead to biased results (Persson & Khamis, 2005). In Chapter 2, we introduce a new approach to analyze TTE data using Bayesian Networks (BNs), (a machine learning algorithm with a semantically meaningful interpretation). We demonstrate the use of BNs model on clinical, gene expression and integrated data with the goal of improving prediction as well as facilitating inference on the structural relationship between variables over time in regards to their contribution to survival outcomes. We demonstrate how to pre-process data in order to analyze it using BNs and applied the new approach to publicly available cancer data from The Cancer Genome Atlas (TCGA) reposi-

tory (Grossman et al., 2016). Our approach exhibit comparable or better performance as compared to other commensurate models for bladder and lung cancer. However, this performance was not observed when the model was used in kidney cancer. A significant benefit from our approach is the flexibility to model and reveal different relationships between phenotypic and genomic characteristics at different time points, improving our ability to understand how factors influence clinical outcomes, and hence, could prove substantial in informing targeted treatment therapies.

In Chapter 3, we review literature materials on predictive modeling in randomized clinical trials (RCTs). RCTs are costly and time-consuming (Morgan et al., 2011; Van Norman, 2016). In this review, we searched for materials on applications of predictive modeling at different stages of RCT. An area of RCTs identified to be substantially influenced by predictive modeling included patient enrolment, where predictive models are used to screen subjects for recruitment into trials as well as using databases to match patients to appropriate trials. Predictive modeling has the potential to mitigate challenges associated with clinical trial failures due to patient accrual. Studies that retrospectively analyzed clinical trial data found that predictive modeling could make clinical trial efficient by identifying subgroup of patients who would have the most benefit from the drugs (Kueffner et al., 2019) for inclusion in the trial which can limit exposure to potentially harmful drugs to non-responders. Other advantages and challenges associated with predictive modeling in RCTs are also discussed.

In Chapter 4, we present a new approach that uses prediction and extends principal stratification (PS) framework (Frangakis & Rubin, 2002) to estimate causal treatment effects in the presence of intercurrent events (ICE). RCTs are prone to post-randomization factors e.g. treatment discontinuation, treatment switch, or adverse events that might influence treatment effects evaluation. This study is motivated by the recent ICH E9/R1 guidelines on estimands and sensitivity analysis in clinical trials that require strategies for addressing intercurrent events to be clearly stated (International Council for Harmonization, 2019). Here, we adopt PS framework where we first use predictive modeling on baseline covariates to predict and allocate subjects to strata based on their counterfactual likelihood of experiencing an ICE under the alternative treatment assignment,

after which we use the appropriate stratum with a homogeneous group of subjects to estimate the causal average treatment effect as well as the weighted average treatment effect. Using simulations, our approach demonstrated a better performance in estimating the causal treatment effect as compared to the standard intent-to-treat (ITT) approach when there is heterogeneity in the impact of the treatment across subjects.

Lastly, we note that there are limitations and challenges associated with the proposed modeling approaches. We discuss these and future directions in each chapter.

Chapter 2

Prognostic Models with Data Integration of Clinical Characteristics and Gene Expression Data Using Bayesian Networks

Abstract

Recent advances in technology have generated vast amounts of heterogeneous types of data collected from human participants with a variety of conditions. Lately, there has been considerable interest in integrating information from multiple sources to better understand patient heterogeneity and help to inform treatment regimens and achieve precision medicine. Time-to-event analysis has wide application in the biomedical domain, especially in cancer studies, with overall survival a common measure of prognosis. The Cox proportional hazards model remains one of the dominant models for the analysis of time-to-event data. However, violations of its assumptions may lead to biased results. Thus, more flexible methods are needed to address patient heterogeneity and violations of the proportional hazards assumption. This study develops survival analysis models using Bayesian Networks that integrate both clinical characteristics and gene expression data for improved inference of prognostic variables and associations among variables. Importantly, it also provides the ability to infer different prognostic relationships at different survival times, which may shed new light on factors that affect prognosis throughout the disease course. Publicly available data from The Cancer Genome Atlas (TCGA) were used to develop and assess the performance of the survival Bayesian Networks. We evaluated model performance using concordance index by comparing our approach to other commensurable strategies as well as standard methods using penalized Cox proportional hazards models. Models using our graphical network approach can infer or confirm meaningful relationships among patient characteristics at multiple timepoints, supporting biomarker discovery and help inform the development of targeted interventions. Despite our emphasis on inference, results from our approach show comparable or better prognostic performance for some cancer types, which suggests that the inferred relationships are ones that should

be investigated further for their therapeutic potential and that our approach may provide a useful way to study complex relationships associated with prognosis.

2.1 Introduction

Innovations in genomic data collection have allowed for vast amounts of disparate types of data to be collected from individual subjects, although effective modeling of integrated data is still an area of active research (Hamid et al., 2009; Neums et al., 2020; Zarayeneh et al., 2017). These data types range from patient-reported outcomes and clinical characteristics to high-dimensional molecular data. In the quest for precision medicine, there is a need to harness and incorporate all available and relevant information from these diverse sources in order to capture patient heterogeneity for improved disease progress tracking and treatment regimens. Currently, most prognostic modeling techniques have a pre-defined structure that associates time-to-event outcomes to observed subject characteristics; that is, the outcome of interest is defined *a priori* such that the covariates influence the outcome marginally or through interaction. This, in turn, may be far from the true unknown causal inter-relationship between the covariates as well as their association with the survival outcome. Moreover, inference based on these models is affected by missing covariate values. Of interest is understanding and assessing structural covariate relationships at different time points, which can improve comprehension of the dynamic system related to a patient's journey. In this study, we seek to develop survival prognostic models using Bayesian Networks (BNs) that integrate clinical characteristics and gene expression data for patients with cancer. Here, we focus on cancer prognostic models, although we note that our approach is equally applicable in other health and disease processes.

Machine learning (ML) algorithms have become ubiquitous in the last decade, with many studies showing the potential for more flexible methods to improve performance in regards to diagnostic, prognostic, and predictive models (Gupta et al., 2011; Kim et al., 2019; Kourou et al., 2015; Weng et al., 2017). However, the complexity and adaptability of flexible machine learning algorithms are an obstacle to implementation in a biomedical domain where interpretability is of

primary importance, despite their considerable predictive performance.

A person is a dynamic biological system with a complex network of interactions. For instance, genes function within biological pathways and interact within biological networks, with each interaction playing different roles. Understanding the relationships in biological systems' can play a critical role in informing targeted patient treatment regimens. Network analysis provides a framework to comprehend the underlying structural relationships among predictors, from gene regulatory networks to protein-protein interactions and phenotype-genotype associations (Mulder et al., 2014; Schadt, 2009), with many studies linking changes in gene-protein networks to human disease (Man et al., 2019; Raj et al., 2017; Wang et al., 2009). Information obtained from these networks improves our ability to find the causes of complex diseases (Chuang et al., 2007). As such, informed decisions based on relevant multi-source data may lead to patients reaping the full benefits from targeted interventions, founded on empirically discovered relationships.

The rate of generation of diverse data has outpaced our ability to integrate these multi-platform data into effective models. Many approaches for data integration have been proposed (Neums et al., 2020; Pittman et al., 2004; Thompson et al., 2018; Zarayeneh et al., 2017). These methods have encompassed two major techniques, meta-dimensional and multi-stage modeling (Holzinger & Ritchie, 2012). Meta-dimensional data integration involves simultaneously combining different data during analysis. However, it is challenging when there are structural differences across data types to be integrated. To mitigate this challenge, multi-stage integration techniques have been used where analysis involves processing the data in stages during the model building and analysis process (Ritchie et al., 2015a). However, in multi-stage modeling, information may be lost during the sequential steps of modeling. Hence, both methods have limitations regarding integration. A review of emerging approaches for integrating genomic and phenotype characteristics have been explored (Hamid et al., 2009; Lin & Lane, 2017; Ritchie et al., 2015a; Sun & Hu, 2016). Other studies have integrated multi-omic, clinical, demographic, imaging and epidemiological data to develop diagnostic/prognostic models (Sun & Hu, 2016; Zhu et al., 2017).

This study centers on developing prognostic models using Bayesian Networks (Pearl, 2014) by

adopting a semi-supervised approach to perform survival analysis whilst integrating multi-platform data. Our graphical approach provides a flexible model with good interpretability. We assess the performance of the proposed method using readily available cancer data from The Cancer Genome Atlas (TCGA) repository. We demonstrate the possibility of using this approach to build prognostic models that integrate baseline clinical characteristics and gene expression data to calculate periodic survival probabilities. Although we focus on survival probabilities, Bayesian Networks can integrate data to reveal relationships that are not dependent on a single outcome, providing an interesting new perspective in the data integration space. Furthermore, by leveraging Bayesian Networks, we show that one can learn different structural relationships at different periods, which we believe could be critical in understanding biological systems exhibiting dynamic regulatory relationships.

2.2 Materials & Methods

2.2.1 Bayesian Networks

Our proposed approach is based on Bayesian Networks (BNs), also known as belief networks. BNs are a special class of graphical models that encode conditional relationships between variables via Directed Acyclic Graphs (DAGs) (Pearl, 2014). DAGs are structural graphs that have directed edges with no cycles. BNs have two main components: a qualitative and a quantitative part. The qualitative component corresponds to the network structure which defines the dependencies between variables (nodes/vertices). Let the network structure of a BN can be represented by $G = (X, E)$ where X is the set of variables or nodes in the network and E, the directed edges connecting the variables which reflect the conditional dependencies between the connected variables. The quantitative component of a BN is the respective set of parameters quantifying the marginal and conditional probability distribution of the variables in the model. Hence, we can completely represent a BN by $BN = (X : \Theta, E)$ with Θ representing the set of marginal and conditional parameter distributions.

Consider a set of variables $X_1, \ldots, X_N$ with the corresponding networks shown in *Figure 2.1*

(N=3). Here, the parent-child relationships can loosely be interpreted as an independent-dependent variable relationship, although with a conservative implication of association rather than causation in an empirically learned network structure. Dependency in a BN is explained through a principle known as *d-separation*, which is used to infer the expected pattern of dependencies given any pattern of paths in the model (Pearl, 2014). That is, two variables are *d-separated* if changing the status of a third variable influences the dependency status between the two variables. Divergent and serial connections have *d-separated* paths (*Figure 2.1*). With *d-separation*, we can learn which variables are independent of each other given a third set. A set of parents, children, and children's parents (spouses) of a node/variable is referred to as a Markov blanket (MB). An important property of an MB regarding inference from our proposed modeling approach is that a variable is conditionally independent of all other variables given its Markov blanket (Witten et al., 2017).

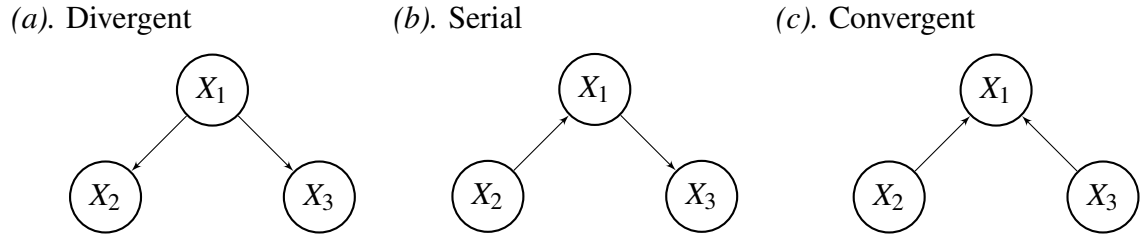


Figure 2.1: **Simple Bayesian Networks**: In (a) & (b), X_2 and X_3 are marginally dependent (or conditionally independent) while for (c), X_2 and X_3 are marginally independent (or conditionally dependent)

We express the joint distribution of the BN in *Figure 2.1.a* probabilistically as

$$Pr(X_1, X_2, X_3) = Pr(X_1) \times Pr(X_2|X_1) \times Pr(X_3|X_1) \tag{2.1}$$

Note that the consequent probabilities from the joint distribution of the variables are recursively decomposed using chain rule. If we consider θ as parameters associated with the probability distribution of a node, then, in general, a BN with p variates is expressed using

$$Pr(X_1, ..., X_p|G, \Theta) = \prod_{i=1}^{p} Pr\left(X_i|X_{(i)parents}, \Theta_{X_i}\right) \tag{2.2}$$

where G represents the network structure and Θ the marginal or the set of parameters corresponding to the probabilities of each random variable given its parent node-set. All parameters of a BN have a semantically meaningful interpretation as opposed to black-box models like multilayer neural networks. Data-driven modeling of a BN involves learning the structure and the parameters of the model. For this study, the `bnlearn package` (Scutari & Ness, 2012) in R is used to learn the BN, where Hill-Climbing (HC) algorithm is used to learn the structure. HC uses a greedy search technique to identify a DAG that best fits the data by evaluating different DAGs, iteratively adding/removing/reversing possible arcs (Russell & Norvig, 2010), with evaluation of the goodness of fit using Bayesian Information Criterion (BIC).

With both discrete and continuous variable types in the data, as in most cases, conditional linear gaussian distributional assumptions are used (Lauritzen & Wermuth, 1989). One current limitation of this approach is that discrete features are assumed to be categorical with no order and the continuous feature can only depend on discrete features but not vice versa. Continuous features conditional on the respective discrete parents are modeled as sets of standard linear regression models in which the continuous parents are considered explanatory variables.

2.2.2 Survival Analysis

Time-to-event data commonly have outcomes with two main components: *i)* whether the event occurred (status), and *ii)* when the event status was evaluated (time, which represents the event time or time at which status was assessed). In the absence of the event, the time typically represents the time at the last observation (though more generally represents a boundary for when the event did or would have occurred). To cope with these divergent meanings of event time (typically called censoring), special analysis techniques have been developed to handle time-to-event or censored data. The most used approach in the biomedical literature is the Cox proportional hazards model (Cox, 1972) in the context of right-censored data (i.e., either the actual event time or a lower bound indicated a time after which the event would have occurred). The hazard in the Cox PH regression models measures the instantaneous risk of an event at a specific time point conditional on survival

to that time, but what is actually estimated is the ratio of this hazard to the hazard in observations without the same value of a given variable. In the presence of covariates, standard Cox models assume a constant effect of covariates on hazards function over time which may lead to biased results when this assumption is violated (Persson & Khamis, 2005).

There is a limited number of machine learning models that are available to perform time-to-event analysis, which is in part is attributable to handling of censored outcomes. New methods have augmented Cox models with neural networks (Ching et al., 2018; Katzman et al., 2018) while others have considered converting time-to-event into a classification problem where censored subjects are treated as event-free (Štajduhar et al., 2009). Another study considered using inverse probability of censoring weights to account for censoring when developing Bayesian Networks (Bandyopadhyay et al., 2015). Our approach takes a different route on modeling the data which involves discretization of survival outcomes through artificial interval censoring as outlined in the next section.

In order to model the data adopting our approach, we illustrate in the following section how the time to event together with censorship status can be discretized into multiple responses analogous to multivariate data. With the discrete responses, any classification algorithm can be used to perform the analysis. However, the BNs approach is advantageous in this setting, due to its flexibility in learning the structure of the network as well as revealing the different relationships that exist at different periods. Independent modeling of the outcome in each period may result in different structural relationships between the factors being studied across the periods leading to changes in how they influence survival outcome over time. Additionally, once a BN is learned, prediction can be performed even with missing covariate values for new data, which is an extremely common challenge with biomedical data, and one that can be further accentuated for multi-platform data.

2.2.3 Survival Bayesian Networks

Our modeling approach involves interval-based discretization of survival times based on event time and event status. Assuming K periods/intervals, the probability of survival beyond period k is then

obtained from the product of conditional periodic survival probabilities, intrinsically,

$$P(S_{k+1}) = \prod_{j=0}^{k} P(S_j) \tag{2.3}$$

where S is a binary variable representing survival outcome. This is akin to a Kaplan-Meier estimate of a survival function. In this framework, we assume $P(S_0) = 1$, meaning each subject has a survival probability equivalent of 1 before the first period, which is logical since follow up is impractical if the subject was unavailable to start at baseline. Furthermore, we assume there are no competing risks, that is, only one type of event is of interest during follow-up. This results in binary outcomes within each period. This is analogous to a special case of K-multivariate binary outcomes. The outcomes are then analyzed sequentially assuming independence with the probability of surviving beyond period k obtained from the product of survival up to and including period k, $(k = 1,2,3,\ldots,K)$. Essentially, $P(S_{k+1}) = P(S_1) \times P(S_2) \times \cdots \times P(S_k)$ (*Equation 2.3*), where $P(S_1)$ is the proportion of subjects who do not experience the event during the first period (survival past the first period). $P(S_2)$ is the proportion of subjects who do not experience the event in the second period given that they neither experienced the event nor were artificially censored during the first period, and so forth. Notice that starting with p_2 up to p_K, we only consider subjects who do not experience the event in the preceding period.

As an illustration, consider a simple case of time-to-event right-censored data for 10 subjects as shown in *Table 2.1*. Columns 2 and 3 are a representation of typical survival data. Column 4 (*Status Period*) represents the period during which an activity (*event/censor*) was observed on the subject based on pre-defined intervals (1-year intervals). In addition, for censored subjects, we proceed to codify the subjects' *Status Period* as follows: If the subject's actual time of censoring is before the mid of the pre-defined interval, then we *artificially censor* the subject in the prior period (*subject 4*), otherwise they are artificially censored in the current period (*subjects 2, 8 & 9*). Discretized survival outcomes are then created (*columns 5:11*) based on the *Status Period* column, and the actual *Status* column as a sequential indicator of the binary event. Generally, the process of creating new binary responses only requires the knowledge of event time and the status at event

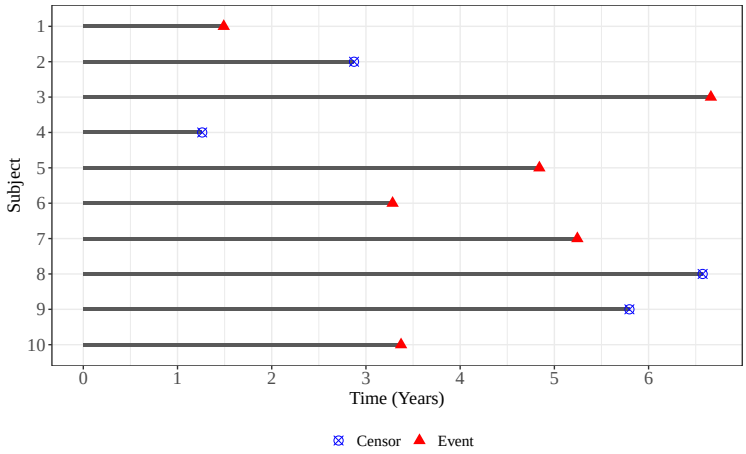
time. If the study event (e.g. death) occurs at time t where t falls in period k, then all the responses preceding period k (i.e $1,2,\ldots,k-1$) are assigned values 0 (event-free), and the status at period k is assigned value 1 (event). On the other hand, if a subject is censored at time t which falls in period k, then, a rule-based artificial censoring is used to create the responses, such that, if the subjects' time of censoring occurred pre-midpoint of period k interval, then the subject is assigned value 0 (artificial censor/event-free) at period $k-1$ and assigned a missing status for the remainder of the subsequent periods. Similarly, if the subjects' time of censoring occurred post-midpoint of period k interval, then the subject is assigned value 0 (artificial censor) at period k and the subsequent periods assigned missing status. These can be visualized in *Figure 2.2*. The new columns, "*Status* P_k", then serve as the new responses and are modeled independently with a subject's survival probabilities calculated sequentially from the models.

Table 2.1: Discretization illustration data

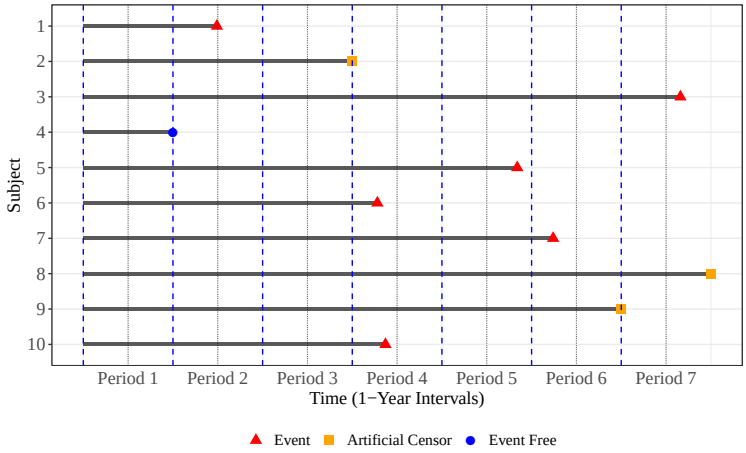
			Pre-Defined Intervals	Discretized Survival Outcomes						
Subject ID	Time (Years)	Status	Status Period	Status P_1	Status P_2	Status P_3	Status P_4	Status P_5	Status P_6	Status P_7
1	1.49	1	2	0	1					
2	2.87	0	3	0	0	0				
3	6.66	1	7	0	0	0	0	0	0	1
4	1.26	0	1	0						
5	4.84	1	5	0	0	0	0	1		
6	3.28	1	4	0	0	0	1			
7	5.24	1	6	0	0	0	0	0	1	
8	6.57	0	7	0	0	0	0	0	0	0
9	5.8	0	6	0	0	0	0	0	0	
10	3.37	1	4	0	0	0	1			

In the presence of covariates, K independent Bayesian Networks are learned for each period with inference performed on the binary responses. During the BN structure learning process, we impose the constraint that all the variables be associated with the binary response. This is a characteristic of most standard modeling approaches where the outcome of interest is always modeled as a function of covariates, e.g. classification/regression methods. Constraining all the other variables in the network to have an association with the binary response allows for prediction of the outcome with all the variables having an influence. This leverages the Markov Blanket property of a Bayesian Network.

We then perform prediction on the binary status variable using the observed covariate values



(a) Right-censored visualization



(b) Discretized right-censored visualization

Figure 2.2: Visualization of right-censored time-to-event data and discretized survival times.

as evidence. That is

$$P(S|E) = \frac{P(S,E)}{P(E)} \tag{2.4}$$

where S is the binary status variable and E is the available evidence. For instance, considering the interest in the probability of no event for an individual in each period, i.e. $S = 0$, we query the learned BN using

$$P(S = 0|X_{observed}) = \frac{P(S = 0 \cap X_{observed})}{P(X_{observed})} \tag{2.5}$$

Our approach considers independent modeling of outcomes in each time period but this can be extended to consider periodic structural dependency as in the case of dynamic BN approach.

However, subject characteristics are only available at baseline limiting applicability of dynamic BN or other approaches such as Hidden Markov models, otherwise, for longitudinal cases, a dynamic BNs can be adopted (Exarchos et al., 2014), or Hidden Markov models depending on the context.

2.2.4 Analysis Framework and Inference

Both clinical and gene expression data are preprocessed before learning the structure and the parameters of the BN. Analysis were performed on subjects with both clinical characteristics and gene expression (RNA-Seq) data available. In addition, binary responses based-on survival time and status variable were created using one-year intervals to define the periods.

We then performed 5-fold cross-validation where folds were created randomly while trying to maintain a balanced proportion of events in each period across the 5 folds. With high dimensional data, learning the structure of the Bayesian Network is a challenging task due to the number of covariates and the computational resource cost problem associated with increasing number of nodes. Therefore, on the gene expression training set, we subjectively selected 10 genes based on p-value and effect size (β coefficient) from the univariable Cox regression model on each gene prior to modeling using BNs (being careful to perform this selection only on the training data for each fold). This involved ranking genes based on p-value followed by their absolute effect size, then arbitrarily selecting the top 10 genes. For the binary periodic responses, BNs and cross-validated logistic regression models were fit on the training set (*Figure 2.3*).

Following gene selection, and within each periodic interval, the structure of the BN is learned by bootstrapping different network structures using the HC algorithm with the final structure of the network being the average of the learned structures. This is analogous to an averaged ensemble of networks, where the averaging on the structure entails using both directionality of influence and strength of relationship between pairs of variables. We also impose a relation constrained such that all variables in the network be associated with the response (binary periodic response). Once the structure of the network is learned, the parameters of the network are obtained from the averaged network using maximum likelihood estimation.

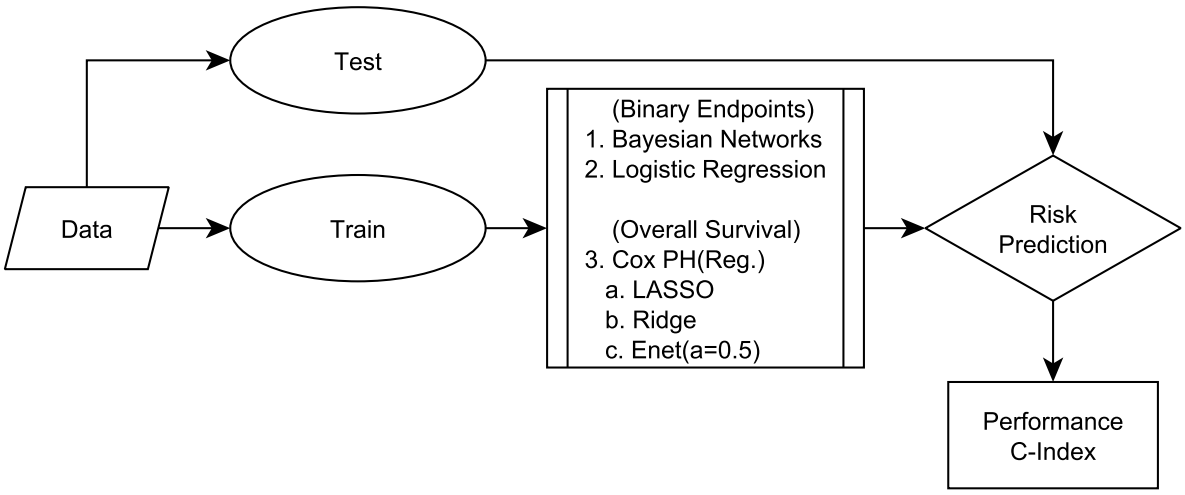


Figure 2.3: Analysis flowchart.

With the fitted BN model, prediction on new data (testing set) is conducted with new data as evidence and probability of survival in each period, $P(S = 0)$, as event. With this framework, prediction of survival during each period is possible for all subjects regardless of censoring. BN learning and prediction is carried out independently for each of the periods, while allowing for change in structural variable relationships across the periods. However, the final inference on subject survival over time is based on the product of survival during each individual period.

Since we were also interested in the performance of the standard approaches, Regularized Cox proportional hazards regression models with α weighting parameter were trained (*Figure 2.3*). That is, ridge $\alpha = 0$, LASSO $\alpha = 1$ and Elastic Net $\alpha = 0.5$ regularization (glmnet definition). These were conducted on the same training set with unfiltered genes but with overall survival as the main outcome, rather than the transformed responses. The trained regularized Cox models were then used to predict the risk score for each subject at the specific time points corresponding to the interval boundaries from the initially prespecified discretization time point cutoffs. Even though we caution on the comparability of the two modeling approaches, we wanted to determine how standard approaches performed on the same data.

The analysis was executed independently on clinical characteristics, gene expression, and the integrated data (both clinical characteristics and gene expression). To assess the performance of the

models, we used the concordance index (c-index) (Pencina & D'Agostino, 2004). C-index is the probability that for two randomly chosen subjects, the subject with higher predicted risk value will experience the event first. The c-index is analogous to the ROC-AUC, and it measures how well the model ranks two random individuals in terms of survival and does not depend on the choice of time for evaluation of the model.

The above is repeated with random resampling of the data herein referred to as different seeds for assessment of the stability of the models. Here, resampling of the data to generate the training and testing sets was performed 100 times. Evaluation of the model performance was then carried out on the c-indices from the multiple seeds.

2.3 Application to Cancer Data

2.3.1 Cancer data profiles

Publicly available data from The Cancer Genome Atlas (TCGA) were utilized in this study. The cancer types selected for the study included bladder, kidney and lung cancer. Clinical characteristics and gene expression data were downloaded through the Genomic Data Commons Data Portal (Grossman et al., 2016). Baseline clinical characteristics selected were based on known cancer-specific prognostic factors and whether there was considerable availability of values for most of the subjects. The selected variables included age, sex, tumor stage, race, diagnosis subtype and status at last follow up *Table 2.2*. Preprocessing of RNA-seq data involved normalization of raw gene counts using `edgeR` (Robinson et al., 2010) and `Limma` (voom function) (Ritchie et al., 2015b) packages in R. Genes with low variance were then filtered out using median absolute deviation (MAD), which is more robust and resilient to outliers than standard deviation. To remain with approximately 12,000 genes, we used MAD thresholds of 1.4, 1.3, 1.34 for bladder, kidney, and lung cancer respectively.

The tumor stage for bladder cancer was reclassified into low and high-risk corresponding to the combination of Stage I & Stage II and Stage III & Stage IV respectively. This was due to the low counts on the number of subjects with Stage I cancer (n=1). For lung and kidney cancer, we

Table 2.2: Baseline demographic profiles for each cancer type

Characteristics	**Cancer Type**		
	Bladder (N=335)	Lung (N=454)	Kidney (N=521)
Age, Median[Q1, Q3]	69[61, 76]	66[59, 72]	61[52, 70]
Sex, n(%)			
Male	91(27.2)	204(44.9)	339(65.1)
Female	244(72.8)	250(55.1)	172(34.9)
Tumor Stage, n(%)			
Stage I	1 (0.3)	250(55.1)	261 (50.1)
Stage II	96(28.7)	106(23.3)	56 (10.7)
Stage III	119(35.5)	74(16.3)	122(23.4)
Stage IV	119 (35.5)	24(5.3)	82(15.7)
Race, n(%)*			
White	280(86.7)	347(86.3)	453(88.1)
Black	20(6.2)	47(11.7)	53(10.3)
Other	23(7.1)	8(1.9)	8(1.6)
Diagnosis Subtype, n(%)*			
Papillary	68(20.6)	-	-
Non-Papillary	262(79.4)	-	-
Follow up vital status, n(%)			
Alive	180(53.7)	282(62.1)	349(67.0)
Dead	155(46.3)	172(37.9)	172(33.0)

* Some subjects missing values; the numbers in parenthesis represent percentages

only considered age, sex and tumor stage as clinical characteristics to include in developing the prognostic models, whereas for bladder cancer, age, sex, race, and diagnosis subtype were used.

2.3.2 Discretizing survival time

For all the cancer types, we used one-year intervals of survival to discretize the periodic responses. These intervals were defined as (0-1], (1-2], ... years with the last periodic interval chosen such that the remaining at risk proportions of subjects was at least 10 percent of the sample study total. Based on these criteria, we ended up with 5 periods for both bladder and lung cancer while kidney cancer has 8 periods. *Table 2.3* shows the distribution of binary outcomes across the periodic intervals. The censored column represents subjects who were artificially censored, while the at-risk column represents the total number of subjects who are at risk of death or censor during that period. Note

that for bladder and lung cancer, the survival time is partitioned into 5 periods corresponding to yearly periodic survival with the last period being survival beyond the 4^{th} year from the day of diagnosis. Similarly, the last survival period for kidney cancer implies survival beyond year 7.

Table 2.3: Cumulative summary of discretized survival endpoints

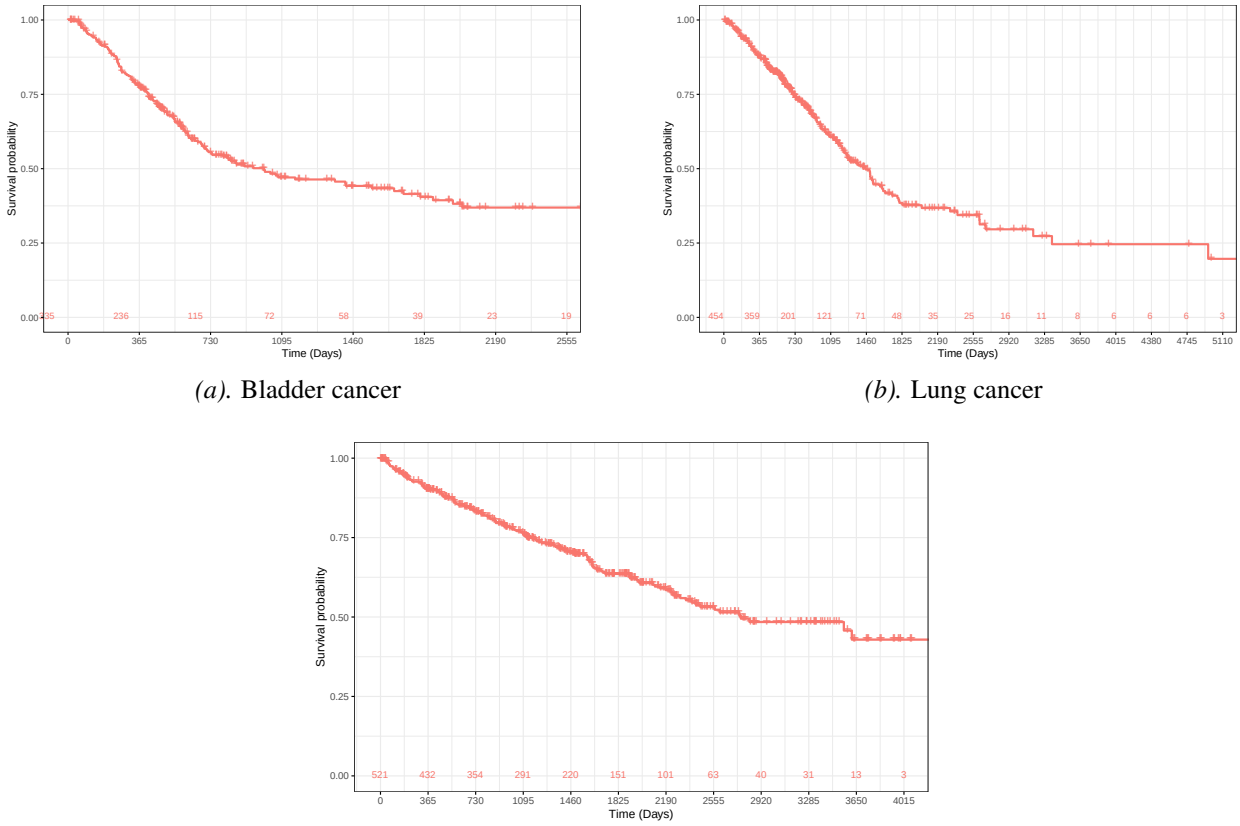
Status Time (1-Year Intervals)	**Cancer Type**								
	Bladder (N=335)			Lung (N=454)			Kidney (N=521)		
	Censored	Died (n=155)	At-Risk (% of total)	Censored	Died (n=172)	At-Risk (% of total)	Censored	Died (n=172)	At-Risk (% of total)
Period 1	29(16.1)	71(45.8)	335(100)	44(15.6)	52(30.2)	454(100)	42(12.0)	49(28.5)	521(100)
Period 2	65(36.1)	55(35.5)	235(70.1)	113(40.1)	45(26.2)	358(78.9)	45(12.9)	31(18.0)	430(82.5)
Period 3	28(15.6)	15(9.7)	115(34.3)	48(17.0)	31(18.0)	200(44.1)	35(10.0)	28(16.3)	354(68.0)
Period 4	11(6.1)	4(2.6)	72(21.5)	31(11.0)	19(11.0)	121(26.7)	51(14.6)	20(11.6)	291(55.9)
Period 5	47(26.1)	10(6.5)	57(17.0)	46(16.3)	25(15.4)	71(15.7)	50(14.3)	19(11.0)	220(42.2)
Period 6	-	-	-	-	-	-	41(11.7)	10(5.8)	151(29.0)
Period 7	-	-	-	-	-	-	29(8.3)	8(4.7)	100(19.0)
Period 8	-	-	-	-	-	-	56(16.0)	7(4.1)	63(12.0)

2.4 Results

Preliminary assessment of overall survival is performed using Kaplan Meier (KM) curves for the three cancer types, (*Figure 2.4*). We observe a rapid decrease in the curve for bladder and lung cancer reaching median survival in approximately 2.5 years and 4 years respectively. On the other hand, the rate of decrease in the KM curve for kidney cancer is gradual, reaching the median survival time at about 7.5 years.

Using univariate cox models to empirically select prognostic genes, a summary of the 10 most frequently selected genes from the random resampling of the data, (multiple seeds), for each cancer type are shown in *Table 2.4*. For instance, *GSDMB*, *PITX3*, and *FIRRE* were the most frequently selected genes for bladder, lung and kidney cancer respectively.

From a single seed and fold, BNs for clinical characteristics, gene expression data, and integrated data in association with binary periodic responses, ($StatusP_k$), are learned and we provide a sample of the integrated data BNs in the first period for the 3 cancer types (*Figure 2.5*). Since we imposed the restriction that all variables in the network be associated with the response variable,



(a). Bladder cancer

(b). Lung cancer

(c). Kidney cancer

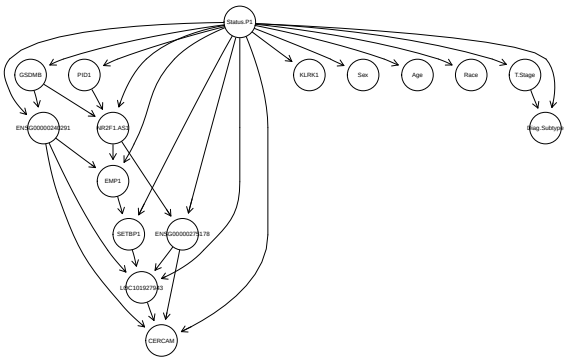
Figure 2.4: **Overall survival Kaplan-Meier curves**: The inset numbers represent the number of subjects at risk at the corresponding specific time.

Table 2.4: Proportion of frequently selected genes (univariate Cox models)

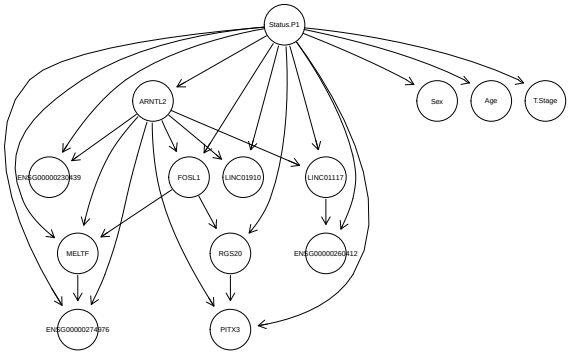
Overall frequency of gene selection					
Bladder		Lung		Kidney	
Gene	Selection Proportions	Gene	Selection Proportions	Gene	Selection Proportions
GSDMB	1.00	PITX3	0.99	SLC16A12	1.00
KLRK1	0.98	ANLN	0.79	COL7A1	1.00
EMP1	0.94	ESYT3	0.60	ZIC2	0.95
ENSG00000240291	0.59	ENSG00000260412	0.56	WDR72	0.84
SETBP1	0.56	MELTF	0.56	FIRRE	0.68
ENSG00000275178	0.40	ENSG00000233609	0.45	OTX1	0.57
TNFRSF14.AS1	0.37	ENSG00000255325	0.40	EMCN	0.53
LAMA2	0.34	LINC01117	0.34	CAVIN2	0.44
NR2F1.AS1	0.29	CLEC17A	0.29	EDNRB	0.41
ENSG00000279254	0.27	DLGAP5	0.27	GPR78	0.29

and as a result of these models being DAG, cycles in the structure are restricted. The network structure exhibited is an average of multiple (1000) bootstrapped network structures and it is ap-

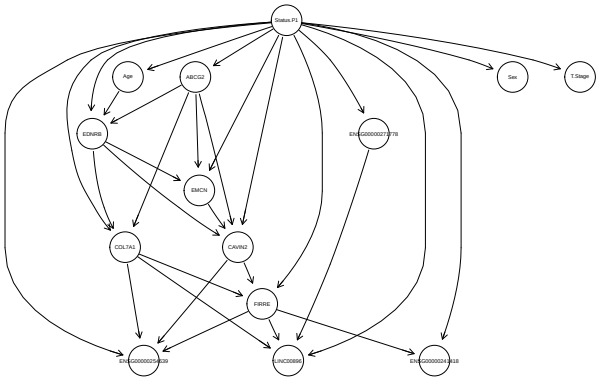
parent that the response variable is the parent of all variables due to the outcome variable type being discrete and the imposed association constraints.



(a). Bladder cancer



(b). Lung cancer

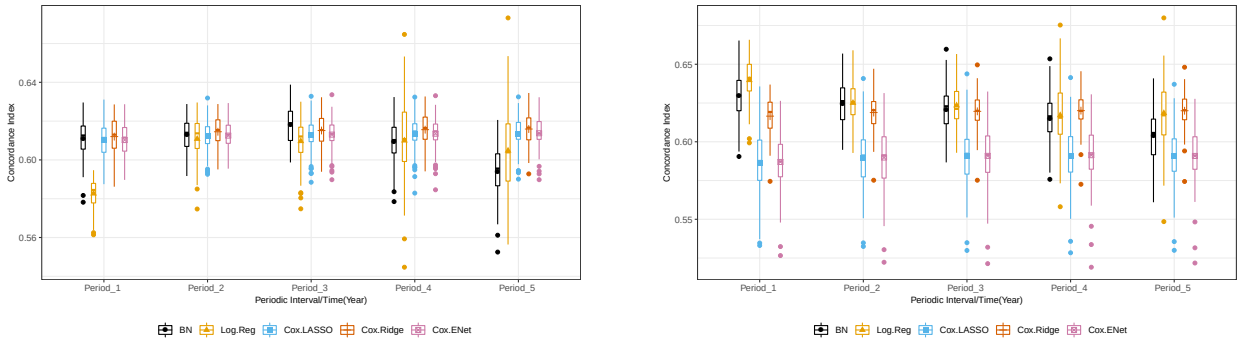


(c). Kidney cancer

Figure 2.5: A sample of learned Bayesian Networks created for the first period for each cancer type.

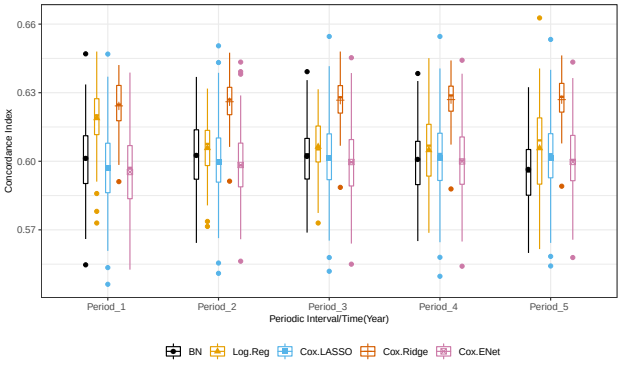
Using the selected networks for the first time period, see *Figure 2.5*, we observe no inter-relationship between clinical characteristics and genes for bladder and lung cancer. However, an association between age and a gene (*EDNRB*), is observed for kidney cancer. From the learned networks, we also identify hub genes (genes exhibiting high connectivity with other genes) that have a potentially direct or indirect impact on other genes as well as the binary survival outcome. For instance, from the selected networks, genes with the most edges as compared to other genes in the networks include *NR2F1.AS1*, *ENSG00000240291*, and *LOC101927943* for bladder cancer (*Figure 2.5.a*); *ARNTL2*, *FOSL1* and *MELTF* for lung cancer (*Figure 2.5.b*). Most genes in kidney cancer are connected to at least two other genes (*Figure 2.5.c*). To evaluate the performance of the

models, we plotted the c-index of the BNs and logistic regression during each period (*Figures 2.6, 2.7, 2.8*). For all the cancer types, BNs exhibited a better or comparable performance on clinical characteristics data as compared to logistic regression. This is also apparent in the integrated data except for kidney cancer. However, performance on gene expression was inferior in all cases. We also included the results from the standard penalized Cox models where the c-index are calculated at the periodic boundaries. Although, and as previously stated, we caution against making direct comparisons between the results obtained from the two approaches. We observe stability in performance across the time points for all the penalized Cox models.



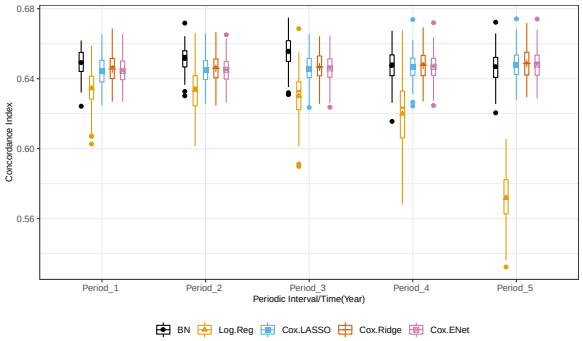
(a). Clinical data

(b). Integrated data

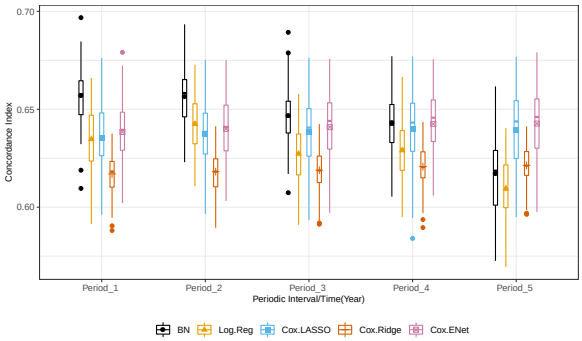


(c). Gene expression data

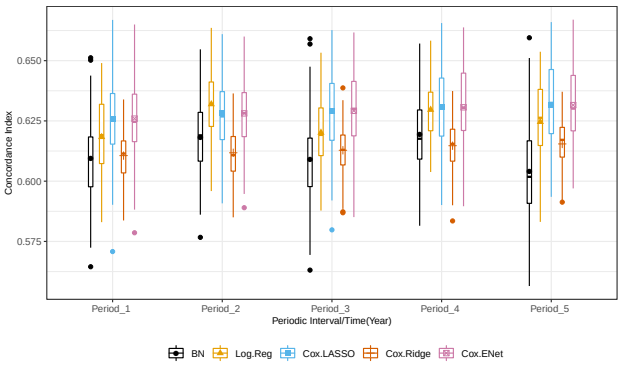
Figure 2.6: Comparison of concordance index for **bladder cancer** cohort with clinical, integrated and gene expression data. BN - Bayesian Networks, Log.Reg - Logistic regression, Cox.LASSO - Cox PH LASSO regularization ($\alpha = 1$), Cox.Ridge - Cox PH Ridge regularization ($\alpha = 0$), Cox.ENet - Cox PH Elastic Net regularization ($\alpha = 0.5$),



(a). Clinical data



(b). Integrated data



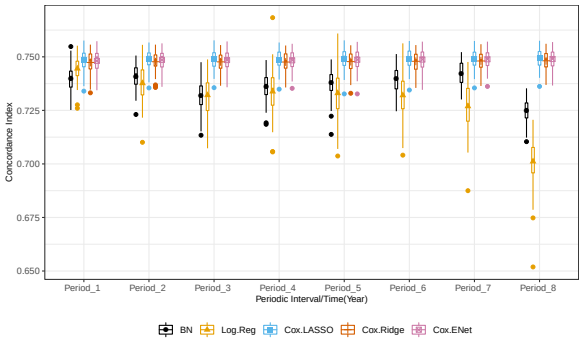
(c). Gene expression data

Figure 2.7: Comparison of concordance index for **lung cancer** cohort with clinical, integrated and gene expression data. BN - Bayesian Networks, Log.Reg - Logistic regression, Cox.LASSO - Cox PH LASSO regularization ($\alpha = 1$), Cox.Ridge - Cox PH Ridge regularization ($\alpha = 0$), Cox.ENet - Cox PH Elastic Net regularization ($\alpha = 0.5$),

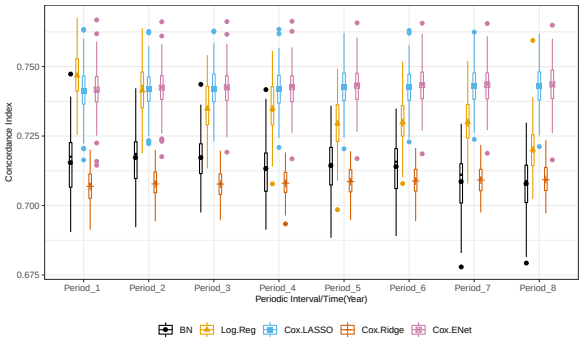
2.5 Discussion

Biological systems have an inherently complex structural association between features. Understanding this phenomenon becomes even more challenging with time-to-event data when learning the structural relationship over time. As such, designing a prognostic model capable of capturing such an association from a cross-sectional data adds complexity to the inference obtained from such models. This study sought to answer two questions: *i*) does creating prognostic models that allow for periodic inference having flexible variable relationships at each period provide useful information, and *ii*) are BN a useful approach to integrating data from two source data?

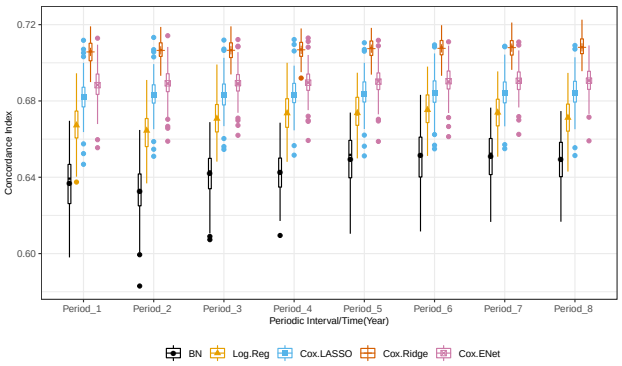
Our approach using BNs achieves better or comparable performance compared to logistic re-



(a). Clinical data



(b). Integrated data



(c). Gene expression data

Figure 2.8: Comparison of concordance index for **kidney cancer** cohort with clinical, integrated and gene expression data. BN - Bayesian Networks, Log.Reg - Logistic regression, Cox.LASSO - Cox PH LASSO regularization ($\alpha = 1$), Cox.Ridge - Cox PH Ridge regularization ($\alpha = 0$), Cox.ENet - Cox PH Elastic Net regularization ($\alpha = 0.5$),

gression for clinical and integrated data as observed in bladder and lung cancer *Figure 2.6* and *Figure 2.7*. This is particularly discernible for lung cancer where there is a pronounced difference in overlaps of the box plots in clinical and integrated data. For kidney cancer, performance on clinical data is comparable, however, with integrated data, we observe that BNs perform poorly (*Figure 2.8*). In all cases, the performance on gene expression is barely comparable and it exhibits fairly poor performance. This is attributable to the problem of overfitting since, with conditional linear gaussian distributions, we assume different distributions for each discrete parent of a continuous child. Thus, given even a moderate number of genes, BNs will tend to overfit. This means the other methods were better able to leverage the gene expression data, and this challenge will

need to be addressed in the future. Even though logistic regression appears to perform better than the BNs in other cases, there are some caveats associated with their performance evaluation. For instance, if the training data has only one class label (events only or non-events only), then a logistic regression model cannot be fit, leading to no predictions made on the test set. This benefits the results obtained from the logistic regression as compared to BNs. Furthermore, there is an increased variability on the estimates obtained from logistic regression compared to BNs.

One benefit from this modeling framework is the potential to reveal complex interactions that might exist between phenotype and genotype characteristics in relation to the outcome of interest. As observed in the BNs created (*Figure 2.5.c*), we note some clinical characteristics (e.g. age) are found to associate with genes (*EDNRB*) which may not be detected if the data types are independently studied. Moreover, the fact that we can allow factors to have different interactions at different time points is manifested in the concept that the biological system works in a dynamic and complex mechanism that varies over time. Thus, we might expect that the expression of different genes at tumor resection, or different clinical characteristics at baseline, might be associated with survival over time. It may be of interest to perform a further investigation on hub genes identified by such networks. Considering bladder cancer hub genes; *NR2F1.AS1*, *ENSG00000240291*, and *LOC101927943* (*Figure 2.5.a*), it is very interesting that all three of these genes are long noncoding RNA genes. In the case of NRF2F1.AS1, it has been previously reported to be associated with oxaliplatin resistance in hepatocellular carcinoma and poor survival in osteosarcoma, potentially by sponging different miRNA (Huang et al., 2018; Li et al., 2019). Possibly, this hints at a relationship among these long noncoding RNA genes in post-transcriptional regulation that was detected by our approach. In the lung cancer data; ARNTL2, FOSL1, and MELTF were hub genes (*Figure 2.5.b*). All have previously been identified as associated with lung cancer survival (Brady et al., 2016; Elangovan et al., 2018; Ma et al., 2020). For example, *FOSL1* is a transcription factor sub-unit that regulates the RAS-ERK cascade, possibly inhibiting apoptosis or driving cell cycle progression (Chang et al., 2003), and *ARBTL2* may facilitate cancer cell metastasis. Furthermore, the ENCODE project identified *ARNTL2* as a possible *FOSL1* target (Consortium, 2011), which

again supports the interpretability of our results. Standard approaches, whether linear or nonlinear, assumes identical inter-relationship of variables across the different time points. Understanding the dynamic variable relationship is crucial in performing targeted interventions.

This approach also allows periodic assessment of risk for individual patients. Conditional on patients' observed disease associated characteristics, having certainty estimates of experiencing an event is imperative if we can create a path for each individual patient on their survival over time which can help inform targeted treatment decisions over time. Moreover, this framework allows for survival prediction on censored observations. In standard analysis, once a patient is censored, we loose information about their future. Of interest for instance would be whether the censorship status could be related to post-censorship factor and this approach provides prediction on censored observations that can reveal meaningful information. In addition, once a BN has been learned, we can query outcomes for new subjects even with limited patient information. If we only have clinical characteristics, we can still get their predicted survival probabilities from the model fit on the integrated data without performing any imputation on the genes.

In this study, we chose the interval bounds for the Cox models to assess their performance as compared to the classification algorithms during the pre-specified periods. For instance, performance was evaluated at time 365, 730,... during periods 1,2,... respectively. There is a limitation with a direct comparison of Cox models with the classification algorithms since they both are estimating different quantities, but it suffices to mention that Cox models are the standard methods and hence were included here for reference. Firstly, Cox models assume constant hazard overtime, which in contrast to our approach, utilizes the entire range of event times when creating models at specific times consequently resulting in stable performance as observed in all the plots. Despite making such noncomparability statements, we still note how closely BNs achieve during the first period in all cases involving clinical characteristics. Moreover, BNs performed poorly on kidney cancer with integrated data, but this trend is also apparent with ridge regularization. This seems to indicate a limited contribution of gene expression in building prognostic models for kidney cancer.

Survival time were split into a pre-defined number of periods using 1-year intervals for all

cancer types for ease of model inference and achieve a systematic framework that captures information about the dynamic biological system. The time intervals were pre-determined based on the context of the study. For instance, different cancer types have dissimilar survival rates, therefore, of interest might be periodic six-month survival or periodic one-year survival for a specific cancer type. Most reported in the literature are 5-year survival rates, however, it is also important to determine survival probabilities for individual subjects at different time periods based on observed characteristics. One could consider extending change point detection techniques to identify such specific times at which the hazard rates change (Goodman et al., 2011) and utilize it for periodic cutoffs.

The assumption of non-informative censoring in time-to-event data motivates the use of artificial censoring. Explicitly, subjects who are censored have the same probability of experiencing the event as subjects who remain in the study. In essence, artificial censoring balances survival estimates by considering subjects who are lost to follow-up prior to the midpoint of the interval as not contributing to the survival estimators for the current period and subjects who survive beyond the midpoint but lost to follow-up before the interval cutoff contributing to survival estimators during the current period.

One limitation of this modeling approach is the stability in performance for the subsequent periods if there is a substantially fewer number of subjects to learn the BNs. Another limitation in our model involves the independent modeling of outcomes periodically rather than a sequential dependency of survival outcomes on their preceding survival estimate. As much as this is a limitation, it offers an advantage over other techniques since feature association can flexibly co-exchange regarding their association at different periods. Future work will consider using dependency on prior state structure when considering predicted probabilities of the current state. Such approaches are directly applicable to panel data where developed methods closely related to BNs include but not limited to Dynamic Bayesian Networks and hidden Markov models. Another future direction will be to consider using an undirected graph to develop prognostic models.

2.6 Conclusions

Bayesian Networks is an invaluable tool that provides a representation of relationships between variables of interest in the situation of uncertainty while handling complexity in the data. The structure of the DAG corresponds to a set of conditional independence assumptions, which can be helpful in revealing and understanding causes of complex diseases. In this study, we have demonstrated how BN can be adopted to model time-to-event data that allows for flexibility in variable relationships over time as well as integrating multi-platform data. In addition, this study shows how we can build an inference tool on time-to-event analysis to understand the association existing between variables overtime which can instigate discussions into developing targeted intervention from the learned relationship.

Chapter 3

Predictive Modeling in Clinical Trials: A Narrative Review

Abstract

Rapid technological advancement has seen increased development of computer algorithms capable of facilitating fast and efficient decision making in clinical trials. This has allowed predictive machine learning algorithms to evolve into being able to model complex systems that were once deemed infeasible, such as in the drug-development sector. Despite these advances, there is still limited progress in the utilization of predictive models in different facets of the highly regulated clinical trial domain. In this article, we review the literature on predictive modeling in clinical trials. We also consider the benefits and challenges of applying these models in a randomized clinical trial setting. Future directions of predictive modeling in the context of clinical trials are discussed.

3.1 Introduction

Predictive modeling involves first modeling outcomes based on input features and a given functional relation (Breiman, 2001) and then utilizing future observable features with the predictive model to predict future outcomes. In a randomized clinical trial (RCT) setting, these input features are typically clinical measures obtained from a subject's health data (e.g. demographic information, vital signs, height, weight, disease duration etc.). These input features are important in many aspects of an RCT from subject screening for inclusion/exclusion, randomization stratification, efficacy/effectiveness covariate adjustments to adverse event (AE) coding (as in treatment-emergent vs non-treatment-emergent AE's). These features are useful in building a predictive model that allows for an efficient clinical trial. While predictive modeling has proven to be effective in other fields, it is imperative to understand the associated benefits and the future of predictive modeling in RCTs.

Predictive modeling is sometimes used interchangeably with machine learning (ML) or predictive analytics. Here, we consider ML as distinct statistical techniques used to construct predictive models, while predictive analytics refers to the commercialized application of predictive modeling which is beyond the scope of this review. In the context of RCTs, predictive modeling and ML are invaluable tools that have the potential to solve complex problems that were considered infeasible in the past (Obermeyer & Emanuel, 2016; Passos & Mwangi, 2018). The past two decades has seen an explosion of novel machine learning algorithms which has been driven, in part, by advances in computing-resource technologies. Additionally, the amount of data collected in the past decade alone is possibly more than what had been collected in all other years combined (Science Daily, 2013). Despite these growths, application of novel ML algorithms has been accompanied by regulation, especially in the clinical trial setting, where interpretability is of greatest importance. Hence, simpler and more interpretable ML algorithms are commonly preferred (e.g. linear regression logistic regression, decision tree, etc.).

Another class of statistical modeling that is closely tied with predictive modeling involves the use of Bayesian methodologies. There are three basic components of Bayesian models; prior, likelihood and posterior. In the RCT setting, the prior component is the information from a previous trial or expert knowledge that can be combined with accruing information from trial data to make a decision in regard to early drug efficacy or futility evaluation enabling an efficient clinical trial. A special category of designs in clinical trials that utilize Bayesian methods (Bayesian adaptive designs) are efficient due to their unique capability with respect to early decision-making to stop the trial for futility or efficacy.

Clinical trials are costly and time-consuming: it takes approximately 10 to 15 years from preclinical testing to the approval of a new drug by regulatory agencies, with the average cost of the process running over $1 billion (Morgan et al., 2011; Van Norman, 2016). Furthermore, about 70% of clinical trials fail due to lack of efficacy or safety with over 55% failing due to inadequate efficacy (Fogel, 2018; Harrison, 2016). Moreover, almost 9 out of 10 new drugs fail in the human testing phase (Van Norman, 2016). Therefore, there is a need to mitigate this challenge of cost and

time while improving clinical trial outcomes. Predictive modeling has the potential to facilitate clinical trial efficiency in areas such as patient recruitment, safety assessment, trial monitoring, and efficacy and effectiveness evaluations.

Recently, there has been an increased interest in utilizing historical data, real-world data (RWD) or in other cases real-world-evidence (RWE) in RCTs (Garrison et al., 2007; Swift et al., 2018). By retrospectively analyzing clinical trial data, prediction models developed from community-wide challenges have demonstrated substantial performance using open clinical trial data (Abdallah et al., 2015; Guinney et al., 2017; Kueffner et al., 2019; Seyednasrollah et al., 2017). For example, the top 10 prediction models in the Prostate Cancer Dream Challenge outperformed the current gold-standard prediction model for prostate cancer survival (Meier et al., 2016) which could prove useful in patient stratification during randomization. Many such promising predictive models have lingered in their development platforms after the conclusion of the challenge that created it, awaiting use in clinical research.

Here, we review the literature on predictive modeling at different stages of RCT conduct and draw attention to the emerging potential of predictive modeling to improve clinical trial efficiency and optimization.

3.2 Methods

In this review, we searched for articles published between 2000 to 2020 in scientific databases including PubMed, ScienceDirect, Web of Science, and Google. Specifically, search terms included, "predictive modeling", "clinical trials" and other gray literature articles on predictive modeling in clinical trials. Most of the research articles identified were either a retrospective evaluation of predictive modeling on clinical trial data or were based on future potential applications of predictive modeling in RCTs. We recognize the closeness of predictive modeling with adaptive design approaches and we supplemented our search to include predictive modeling approaches augmented by adaptive clinical trial designs which have the flexibility to modify one or more aspects of clinical trials based on accumulating information.

3.3 Results

There are several aspects of predictive modeling that could benefit clinical trials. Here we highlight key features of predictive modeling and their associated uses in the various stages of clinical trials (e.g. pre-screening, trial optimization and, safety and efficacy evaluation). We note that at the time of writing this paper there are many areas of clinical trials that predictive modeling can potentially be applied to, and they cannot all be addressed in this narrative.

3.3.1 Setting the Scene: A perfect model

Consider a hypothetical scenario of a clinical trial where there is a *perfect* prediction for some parts of the trial. For instance, these predictions may be in the form of simple tasks such as the number of patient accruals in a specific period, the expected number of safety issues, the population of subjects who will benefit the most from the study drug as well as quantifying the effectiveness of the study drug. Note that a *perfect* prediction model for either of the listed parts of the clinical trial can tremendously improve the efficiency of a clinical trial in terms of cost and time. There is currently no such thing (yet) as a *perfect* model, hence, these examples are too optimistic but represent the "optimum". However, some algorithms that are "non-perfect" but good have been developed.

The goal of a perfect model is to speed up the trial as well as reduce the financial cost associated with clinical trials (Morgan et al., 2011; Van Norman, 2016). Some components of a clinical trial that are targets for predictive modeling utilization are concerned with; *i*) safety and toxicity of an intervention (Menard et al., 2019; Seyednasrollah et al., 2017), *ii*) whether there are subjects who will respond to the study drug (Chekroud et al., 2016; Gullick et al., 2017), *iii*) overall efficacy and effectiveness evaluation of an intervention (Ezzati & Lipton, 2020), and *iv*) developing new information from the collected clinical trial data related to drug dynamics as well as the molecular drug profile that will inform future drug development or drug repurposing (Yella et al., 2018).

The current era of high-dimensional complex data e.g. molecular data as well as data from

different platforms requires prediction models capable of improving understanding of patient outcomes (Hernandez & Zhang, 2017), in addition to assessment of drug safety and efficacy. There has been an increased interest in using real-world data (RWD) as well as real-world evidence (RWE) to streamline health care decision-making (Bate et al., 2016; US Food and Drug Administration, 2018). Regulatory agencies are currently using RWD and RWE for monitoring post-market safety and adverse events to make regulatory decisions (US Food and Drug Administration, 2018). RWD and historical data from closely related studies, provides a great platform for developing an informative predictive model. Models utilizing historical data have been efficiently implemented in designs adopting Bayesian approaches where information is borrowed (Wang et al., 2019).

3.3.2 Trial Design: Pre-trial

Predictive modeling from the clinical trial design perspective entails identification and considerations for more robust and clinically meaningful endpoints. This involves the identification of important factors that are most predictive of trial outcome which is useful in understanding uncertainties that might arise during the clinical trial conduct. For instance, models can be designed as in *in silico* clinical trials which are virtual trials where virtual cohorts of patients are created for testing safety and efficacy of new drugs (Pappalardo et al., 2019) or by relying on previous trials and/or observational studies to identify parameters that are most predictive of trial outcomes to mitigate uncertainties associated with trial expectation (Cui et al., 2014).

An appropriate study population for a study drug can be selected using predictive modeling. Strategies using patient demographics, historical, molecular drug profile among other characteristics have been established for the identification of subjects who have the potential to benefit from the drug. Such strategies that are chosen in advance of a clinical trial include predictive enrichment strategies where subjects are chosen based on their likelihood of responding to drugs (Renfro et al., 2016; US Food and Drug Administration, 2019).

3.3.3 Patient Enrollment

About 80% of clinical trials are unable to meet enrolment goals with over 75% of cancer trials failing to enroll a sufficient number of patients (Fogel, 2018). There has been an uptick on the need to use predictive models for patient recruitment to mitigate this challenge (Barnard et al., 2010), from development of prediction models capable of translating free text information from electronic health records (EHRs) data to support patient recruitment (Thompson et al., 2019), to utilization of prediction models to assess patient eligibility (Kopcke et al., 2013). Besides, platforms have also been developed to perform fast and early pre-screening of potential subjects in databases for target populations and subject accrual feasibility assessments (Mudaranthakam et al., 2018). Ni et al. utilized Natural Language Processing (NLP), Information Extraction (IE), and machine learning techniques to develop an automated clinical trial eligibility prescreening tool which demonstrated a substantial increase in screening efficiency for matching patients to clinical trials (Ni et al., 2015). Over the past decade, several technology start-ups have developed platforms to link patients with appropriate clinical trials that meet patient needs.

Predictive modeling should be useful for patient stratification in regard to their potential response to drugs. For example, in the crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression, predictive models reduced the heterogeneity of patients in a highly complex disease by differentiating subgroups of patients (Kueffner et al., 2019) which shows how such models can be implemented during patient randomization. Predictive models used for patient stratification could be instrumental in screening for patients who are the best fit for the trial as well as limiting harm to potential non-responders.

By adopting predictive modeling techniques for patient accrual, there is a potential to improve clinical trial efficiency during patient recruitment by predicting the likelihood of patient's participation in the clinical trial (Ni et al., 2015). Also, fewer subjects are exposed to unnecessary drugs because subjects who are most likely to respond to the drug are identified for inclusion into the trial (Escudero et al., 2011; Kueffner et al., 2019). Hence, financial cost, logistical, and ethical constraints can be mitigated and live-saving treatments can reach patients faster.

3.3.4 Monitoring Trials

During interim analysis, it is typical that some subjects will have missing outcomes. Missing data may influence clinical trial inferences depending on the missingness patterns and how the missing data are handled (Jakobsen et al., 2017). Typically, subjects will have baseline or prior interim data that can be useful in imputing the missing outcomes. Predictive algorithms are a great tool in such settings where missing outcomes are imputed for subjects that have not reached the intended endpoint but have baseline or prior interim data . In addition, use of predictive modeling during interim has the ability to identify subgroup of treatment responders which could inform potential trial modification (Ballarini et al., 2018). As noted previously, there is an overlap in predictive modeling and adaptive clinical trial designs and we briefly review the benefits of typical adaptive designs.

Adaptive clinical trial designs are a type of trial that allows for modifications on the trial based on accumulating information (Berry, 2004). Of significant importance in adaptive designs is the ability to monitor and predict patient accrual (Gajewski et al., 2008; Jiang et al., 2015) and make early decision rules where trials can be stopped early for success or futility based on predictive probabilities (Berry, 2006; Broglio et al., 2014; Gupta, 2012; Krams et al., 2003). Such models also permit for sample size re-adjustment with accumulating trial data. As such, these models represent a clear example of a current use of predictive models in the clinical trial setting that has been found to have great utility.

3.3.5 End-of-trial

At the conclusion of the trial, missing values are imputed using predictive models similar to imputations performed at interim. To assess efficacy, new approaches have been developed that extend propensity score methods to stratify subjects while borrowing information from historical data using power prior technique (Wang et al., 2019). This is particularly beneficial for single-arm studies, especially in trials with rare diseases. Besides, this can provide some information regarding whether the drug is beneficial to some subgroup of subjects which can inform future clinical

trials (Ballarini et al., 2018; Schnell et al., 2016; Weisberg & Pontes, 2015; Kueffner et al., 2019).

With predictive models, new indications for approved drugs can be assessed for future therapeutic potentials (Gottlieb et al., 2011; Li & Jones, 2012; Napolitano et al., 2013; Yella et al., 2018). After the drug has been approved, there is a need for exploring interactions of the new intervention with other co-administered medications to assess adverse events from drug interactions. Predictive models have been developed to predict drug-drug interactions based on drug characteristics and adverse drug events (Cheng & Zhao, 2014; Page et al., 2012).

3.3.6 Challenges

A major challenge of using predictive modeling is the potential complexity of some predictive algorithms. Although sophisticated predictive models can sometimes have high predictive accuracy, there are limitations associated with their interpretability. In many cases, there is a tradeoff between model accuracy and interpretability (Johansson et al., 2011). The most flexible predictive models, capable of modeling the most complex functions of the data, also require a large sample size to effectively estimate parameters of interest. This is especially true when there is a large number of input features.

Model generalizability is also a concern in adopting developed prediction models from one clinical trial to another. For instance, if borrowing information from unrelated historical data or RWD, prediction models developed from these data may be unreliable and can lead to biased inference if the models are adopted for the task at hand.

Due to the need to protect patient information, it is often challenging to obtain the most relevant data for developing predictive models. Because of this, great effort is expended by researchers on developing models on data that are easier to obtain but that are likely suboptimal in terms of what could be achieved. Future considerations for such situations might require development of predictive modeling algorithms in an encrypted framework, after which parameters or model summary characteristics are the only attributes available on the front end, thus limiting access to private data.

Finally, there are challenges associated with adoption of new techniques especially in a highly regulated field that has a well-established standard of practice requiring more validation on new approaches.

3.4 Discussion

Predictive modeling has the potential to dramatically improve clinical trial efficiency in terms of patient recruitment, safety assessments, trial monitoring, and efficacy assessments. It could facilitate a real-time accurate understanding of drug mechanisms and streamline the development of targeted drugs in regards to personalized medicine.

Currently, there is an under-utilization of predictive modeling in clinical trials. However, there are potentially important benefits that are associated with predictive modeling: fewer patients exposed to potentially harmful drugs and improved resource optimization. Prediction in combination with adaptive designs has the potential to significantly reduce the duration of trials, minimizing their cost as well as allowing important drugs to be available to patients sooner. ML algorithms are becoming an indispensable tool for solving complex problems in nearly all walks of life, and some of the studies that have retrospectively analyzed clinical trial data have shown the potential of prediction models to improve clinical trial efficiency. For instance, subgroups of patients who can benefit from the drug might better be screened using predictive models.

There have been considerable efforts to consider novel validated approaches that have gained popularity in other fields. A Project funded by US FDA is currently being carried out to evaluate whether using RWD and RWE can successfully replicate the outcomes from a pragmatic clinical trial (RCT Duplicate, 2020). The study aims to use RWD to reproduce RCTs and compare findings to predict the results of seven ongoing Phase IV clinical trials. If validated, it could pave the way for drug developers to apply for approval of new indications based on predictive modeling and in silico trials, although this is likely still far off. Drug developers might also be interested in predicting future outcomes of drug approval after phase II in order to assess the logistics of the clinical trial moving forward (DiMasi et al., 2015).

Most machine learning approaches involve the assessment of model performance regarding sensitivity and modeling assumptions. This is beyond the scope of this narrative review. However, we note that predictive models should be assessed for their predictive accuracy. This is typically hard to evaluate in a clinical trial setting given that every clinical trial is unique. However, historical data or RWD may provide a platform for development and validation of prediction models. Before such predictive models are eventually accepted as a standard of practice, robust proof is required to validate with randomized pragmatic trials, so that future implementation of such predictive models results in more efficient and reliable clinical trials.

Finally, we note that predictive modeling algorithms may not necessarily apply to every clinical trial. As such, standard methods and ethical concerns may supersede applicability of predictive modeling techniques. To fully adopt predictive models in clinical trials, domain experts will be required to validate developed predictive models. With advances in computational capabilities, assessment of different scenarios in a clinical trial is possible using simulations and validation of the predictive models will require in-depth review by experts before deployment.

Chapter 4

Estimating Causal Treatment Effects in the Presence of Intercurrent Events: A Bayesian Inference Approach Adopting Principal Stratification with Strata Predictive Covariates

Abstract

Treatment effect in randomized clinical trials is often evaluated after post-randomization intercurrent events such as treatment discontinuation/switch, use of rescue medication or even death. However, without appropriate adjustment for these intercurrent events, the treatment effect estimate is very likely to be subject to bias and therefore misleading since it no longer reflects the treatment causal effect. The recently released ICH E9(R1) guidelines on estimands and sensitivity analysis in clinical trials also emphasizes the importance of this adjustment to ensure the statistical validity and clinical meaningfulness of the estimated treatment effect. To adjust for intercurrent events, we adopt the principal stratification framework where we first predict the latent strata membership based on observed baseline characteristics and then evaluate the causal treatment effect within the appropriate principal stratum. In addition, a weighted treatment effect based on observable stratum specific outcomes is calculated. Since the true causal effects of a treatment is not known in a real setting, we assessed the performance of our approach using simulations and compared our results to the standard ITT approach that does not adjust for intercurrent events. In the presence of intercurrent events, our approach demonstrates a reduction in treatment effect bias compared to ITT analysis using MSE and is more robust to heterogeneity in treatment effects between subjects. In addition, this approach can further inform the design of succeeding phases of a clinical trial regarding screening for inclusion/exclusion through predictions of the potential of subjects to experience an intercurrent event.

4.1 Introduction

In Randomized clinical trials (RCTs), selection bias on baseline covariates that can confound treatment effects are mitigated through randomization. However, RCTs are prone to post randomization confounding factors such as treatment discontinuation due to adverse events or lack of efficacy, use of rescue medication, treatment switch, protocol deviation or even terminal events that preclude observation of the outcome of interest. Treatment effects in clinical trials are normally considered causal effects due to their natural control of selection bias through randomization (Schulz, 1998). However, the aforementioned factors that arise in the course of the clinical trial may influence accurate estimation of the treatment effect estimand Gupta (2011). Despite the inclusion/exclusion criteria at the trial onset serving as powerful tools in screening for potential factors that might lead to trial participants experiencing an intercurrent event (ICE), it is unrealistic to assume that no such events will occur. Analyses that fail to consider intercurrent events may subsequently lead to biased inference on the treatment effect estimand. Standard analysis using the intent-to-treatment (ITT) strategy in presence of intercurrent events are generally conservative to treatment effect estimation because of treatment effect dilution by post-randomization confounders. Thus, intercurrent events may limit the maximum inferential benefits from randomization resulting in treatment effect bias (Gupta, 2011; Zheng et al., 2020). In other words, this limits the accurate estimation of the desired treatment effect and can preclude the conclusion of causality. Hence, these events need to be adjusted for during analysis. However, adjusting for post-treatment variables as if they were baseline confounders can similarly bias the treatment effect estimate (Rosenbaum, 1984). Some strategies for addressing such events have been outlined in the recent ICH E9/R1 guidelines (International Council for Harmonization, 2019) with several studies exploring such strategies (Aroda et al., 2019; Keene, 2019; Ratitch et al., 2020). Nevertheless, there is a need for more robust methods that addresses the new guidelines on ICE whilst obtaining unbiased estimates of treatment effects which may be biased by ICEs.

In this work, we propose a method that not only helps to accurately estimate causal treatment

effects but also provides a framework for using factors observed at baseline to identify subjects who are likely to experience intercurrent events in any follow-up clinical trials. Our proposed approach is an extension of *principal stratification* (PS) framework (Frangakis & Rubin, 2002) to estimating causal treatment effects in the presence of intercurrent events by first utilizing baseline covariates to predict intercurrent event under the control/drug alternative assignment and then estimating the causal treatment effects using the homogeneous group of subjects in the appropriate stratum.

This approach is motivated by the ICH E9/R1 guidelines on estimands and sensitivity analysis in clinical trials (International Council for Harmonization, 2019) and inspired by techniques that addresses treatment noncompliance in RCTs (Feller et al., 2017; Roy et al., 2008). The fundamental property of causal inference is based on the concept of *potential outcomes* framework (Rubin, 1986, 2005). In a two-arm study, this involves imagining a hypothetical outcome scenario under the alternative treatment assignment referred to as a counterfactual outcome, with the difference of outcomes obtained from these two concurrent scenarios resulting in a causal effect interpretation. In the presence of intercurrent events, we adopt the principal stratification framework to first classify subjects into strata according to their potential intercurrent event outcomes. Principal stratum refers to the subgroup of subjects who would have homogeneous outcomes in regards to their joint potential outcomes (Vanderweele, 2011), hence, the resulting effects from the principal stratum are causal effects. Several studies have adopted the principal stratification framework on clinical trial data to address treatment noncompliance (Mattei & Mealli, 2007; Odondi & McNamee, 2013; Sheng et al., 2019), while others have augmented PS and principal score methods for predicting strata (Ding & Lu, 2017; Feller et al., 2017; Funk et al., 2011). PS have been used in other context to identify responders to treatments (Porcher et al., 2019). Our approach provides a simple framework that is applicable to other post-randomization factors beyond treatment non-compliance, and with limited assumptions.

In this study, we assume the desired principal stratum of interest to be the stratum where the subjects do not experience the intercurrent event regardless of treatment assignment. Others have referred to the subjects in such stratum as compliers, and the treatment effects from such stratum as

complier average causal effect (CACE) (Imbens & Rubin, 1997) or local average treatment effect (Imbens & Angrist, 1994). We develop the methodology of evaluating causal treatment effects in the presence of intercurrent events and assess performance of the approach on simulated data under different scenarios.

4.2 Methods

4.2.1 Potential Outcomes

Clinical trials are prone to intercurrent events. These are events that occur post-randomization and are challenging to completely prevent using inclusion/exclusion criteria. In the causal inference perspective, once a subject has been randomized to receive either study drug or control, there are two potential outcomes regarding intercurrent events, that is, the subject will or will not experience an intercurrent event. For a two-arm clinical trial, let Z represent treatment assignment with $Z = 0$ corresponding to placebo and $Z = 1$ corresponding to the study drug, which may interchangeably be referred to as treatment. Further, let D represent a binary indicator of intercurrent event status ($D = 0$: *No Intercurrent Event (NICE)*, and $D = 1$: *Intercurrent Event (ICE)*). For simplicity, we assume that subjects with intercurrent events eventually end up having unobserved endpoint outcomes, although we note that there are many possible ramifications to intercurrent events. Even though we focus on unobserved outcomes after ICE, this approach can also be extended to incorporate scenarios that allows for observable endpoints e.g. treatment switch or use of rescue medication. However, the resultant treatment estimand from these scenarios might have a different interpretation. For now, we defer dealing with complex intercurrent events such as treatment switch or use of rescue medication for future research. Further, we assume that D is a composite variable bearing all information pertaining to an intercurrent event without considering the type of intercurrent event (again, we will consider a more nuanced approach in future work).

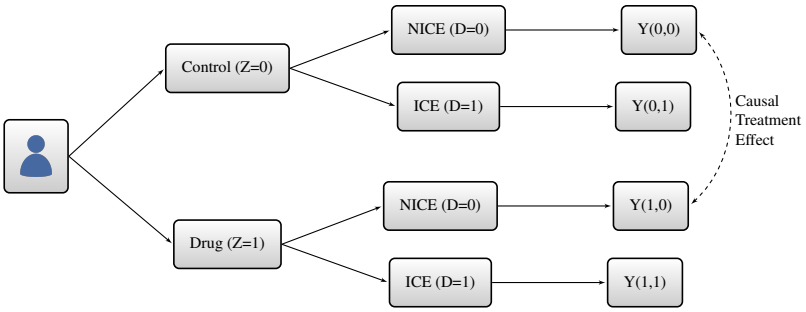
A fundamental property of defining causal effects involves the potential outcomes framework (Rubin, 1974, 1986, 2005). In a two-arm study, each subject i has a set of two potential outcomes, $Y_i(0), Y_i(1)$. In order to conclude causal treatment effect, causal inference assumes that the same

subject is observed under both treatment groups simultaneously; in addition, the subject should not experience an intercurrent event. This is not possible, since we can only observe one outcome at a time for each subject:

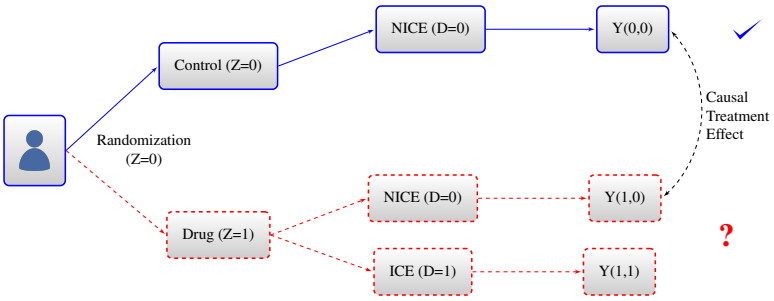
$$Y_i = (1 - Z_i) \times Y_i(0) + Z_i \times Y_i(1) \tag{4.1}$$

where Z is the observed treatment assignment and $Y(0)$ and $Y(1)$ are the associated potential outcomes under both treatment arms. Similarly, once a subject has been randomized to one of the two arms, there are also two potential intercurrent event statuses, $D(Z) = 0$ and $D(Z) = 1$ corresponding to NICE and ICE respectively except that we can only observe intercurrent event status under the assigned arm adding to the complexity of assessing treatment effects. *Figure 4.1.a* shows all the possible paths a subject can take in a randomized clinical trial. The elements of the outcome variable Y, $(Y(z,d))$, correspond to the treatment assignment and intercurrent event status respectively. An estimand that reflects the causal treatment effect would therefore imply comparing $Y(0,0)$ to $Y(1,0)$. This stratum represents homogeneous group of participants who would have similar joint potential outcomes, (NICE in both arms), during follow-up. However, subjects can only be randomized to one group followed by an observed intercurrent event status on the assigned arm, resulting in partially observed information as illustrated in a case of assignment to control with no intercurrent event (*Figure 4.1.b*). Note that the intercurrent event status and the counterfactual outcome under the alternative assignment for this case are unidentifiable. Therefore, the overall goal is to decipher whether the resultant outcomes from the same subject under the alternative treatment assignment would be $Y(1,0)$ or $Y(1,1)$.

One key assumption in causal inference is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980; VanderWeele & Hernan, 2013). Under this assumption, there is only one version of the treatment and no interference between units. That is, the treatment status of a subject does not interfere with the potential outcomes of other subjects (no hidden treatment variation). In addition, this allows for consistency in outcomes under the two arms and the ability to express the outcomes as in *Equation 4.1* (Pearl, 2010). In the absence of intercurrent events, potential outcomes in randomized clinical trials are independent of the treatment assignment. Hence, with independence,



(a) All potential paths for a subject pre-randomization



(b) Observed outcomes and counterfactuals, assuming randomization to control and no intercurrent event

Figure 4.1: Illustration of all paths for a subject in a clinical trial. Y is the final outcome measured on the subject. Elements on Y are treatment assigment z and intercurrent event status D respectively. *NICE- No Intercurrent Event*, *ICE -Intercurrent Event*.

randomization, and the SUTVA assumptions, we can estimate the average causal treatment effect Δ directly from observed outcomes in a two-arm clinical trial. Assuming $\delta_i = Y_i(1) - Y_i(0)$ are individual level causal treatment effects, the average treatment effect is evaluated as

$$\begin{aligned}
\Delta &= E(Y(1) - Y(0)) \\
&= E(Y(1)) - E(Y(0)) && (Independence) \\
&= E(Y(1)|Z=1) - E(Y(0)|Z=0) && (Randomization) \\
&= E(Y|Z=1) - E(Y|Z=0) && (SUTVA)
\end{aligned} \tag{4.2}$$

With the potential outcomes framework, the outcomes $Y(1)$ and $Y(0)$ are fixed and consistent for each subject, hence, the expectations of these outcomes can be expressed independently. If there are no intercurrent events, then causal treatment effect is straightforward to calculate from *Equation 4.2*. On the other hand, the presence of intercurrent events requires further assumptions on estimating causal treatment effects. Subjects are characterized by strata based on their observed

treatment assignment, intercurrent event status and their potential intercurrent event under their alternative treatment assignment. As a result, subjects are considered to emanate from a latent strata mixture.

4.2.2 Principal Stratification and ICE Predictor Covariates

To model the latent strata mixture, we assume there are associated effects of baseline covariates on experiencing an intercurrent event. That is, based on the observed baseline covariates (X), treatment Z, and observed intercurrent event status D, we create a model to obtain the likelihood (probability or propensity) of experiencing the intercurrent event under the current treatment assignment using the latent strata model

$$\pi_{D|Z,X} = P(D|X,Z) = I(Z) \cdot f_X(x) \tag{4.3}$$

where $I(Z)$ is an indicator of treatment assigned taking values 1 or 0 conditioned on the assigned arm Z. For instance, if a subject is assigned to control, then $I(Z) = 1$ for control and $I(Z) = 0$ for drug. This results in two expression/models corresponding to independent modeling of intercurrent event for each arm. $f(x)$ is the standard logistic function with covariates X used in modeling the intercurrent event variable. As a function of covariates, $f(x)$ can be presented as

$$f_X(x) = \frac{exp\left(\delta_{I(Z)}x\right)}{1 + exp\left(\delta_{I(Z)}x\right)} \tag{4.4}$$

where δ's are the arm specific associated parameters obtained using the logit link such that

$$logit\left(\pi_{D|Z,X}\right) = I(Z) \cdot \boldsymbol{\delta}_{I(Z)}\boldsymbol{x} \tag{4.5}$$

As a consequence of having the indicator, two independent models are created for modeling intercurrent event status for each arm. In addition, if the same covariates are used for the two models, the number of parameters to estimate is therefore doubled. The fitted latent strata models are

then used to obtain counterfactual posterior predictive probabilities (propensities) of experiencing an intercurrent event under the alternative assignment.

$$\widetilde{\pi}_{D|Z,X,\delta_{I(Z^*)}} = I(Z^*) \cdot f_{X|\delta_{I(Z)}}(x) \tag{4.6}$$

where Z^* represent the alternative assignment. We refer to the estimates obtained from *Equation 4.6* as the posterior predictive counterfactual probability of an intercurrent event under the alternative assignment. With this approach, all subjects will have the observed intercurrent event status under their true assignment and the predicted probability of having an intercurrent event under the alternative assignment. Allocation of subjects to stratum will then be simulated using MCMC and will be based on these posterior predictive probabilities where a subject's class follows a categorical distribution. A two-arm RCT with binary intercurrent event status generates four strata (*Table 4.1*). With observed treatment assignment and intercurrent event status, a subject can only belong to one of two complementary strata,

$$S \in \{s, s^*\} \ni \begin{cases} S = s & if\ \pi_s \geq \pi_{s^*} \\ S = s^* & if\ \pi_s < \pi_{s^*} \end{cases} \tag{4.7}$$

Here, s and s^* are the complementary strata. For example, a subject randomized to control who doesn't experience an intercurrent event $(D = 0)$ can only belong to stratum 1 or stratum 2 (*Table 4.1*). The elements, π_s and π_{s^*} in *Equation 4.7* represent the counterfactual posterior predictive probabilities of belonging to s and s^* respectively. Therefore, $\pi_{s^*} = 1 - \pi_s$.

Table 4.1: Randomization-Intercurrent event strata distribution

	Strata			
Group	Never(S_1)	AE/Other Event(S_2)	TS/Efficacy (S_3)	Always(S_4)
Control	$D = 0$	$D = 0$	$D = 1$	$D = 1$
Drug	$D = 0$	$D = 1$	$D = 0$	$D = 1$

AE-Adverse Event, TS-Treatment Switch: $D = 0$ - No intercurrent event(NICE), $D = 1$ - Intercurrent event(ICE)

The first stratum, S_1, represents subjects who will not experience the intercurrent event regard-

less of treatment assignment. S_2 represents subjects who will experience an intercurrent event if assigned to the drug but not to the control, whilst S_3 are the subjects who will experience the intercurrent event when assigned to the placebo group but not to the drug. Stratum S_4 represents the group of subjects who will always experience the intercurrent event regardless of treatment assignment. As an illustration, if a subject is randomized to control and doesn't experience an intercurrent event, then this subject is considered to be from a mixture of two potential strata, S_1 or S_2, depending on the likelihood of experiencing the intercurrent event under the drug assignment. If the likelihood of experiencing an intercurrent event *vs* not experiencing the intercurrent event if assigned to drug is low, then the subject is more likely to be allocated to S_1 rather than S_2.

The causal treatment effect of interest is then estimated by comparing the effects in the first stratum, (S_1), which contains a homogeneous group of subjects. In addition, an overall weighted effect of the treatment can also be obtained using the stratum specific effects and model generated latent strata proportions. However, if intercurrent events result in missing outcomes, as in our case, the weighted treatment effect is evaluated only from observable stratum outcomes, hence, some strata (strata 3 & 4 in control, and 2 & 4 in drug) may not be utilized in calculating the weighted treatment effect. For instance, recognize that if ICE result in missing outcomes, all subjects in stratum 4 have missing outcomes, hence, estimates of treatment effect in this stratum will highly be influenced by prior definition.

4.2.3 Estimands

Treatment effects are estimated within each stratum by using an appropriate modeling distribution. For instance, if dealing with continuous outcome, as in our simulated data scenarios, then we can model the outcomes in general using a regression model;

$$Y_{is} = \theta_{0s} Z_{0is} + \theta_{1s} Z_{1is} + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \varepsilon_s \tag{4.8}$$

where θ_{0s} and θ_{1s} are the effects in the control and drug group respectively in stratum s with the Z's being a column vector of 1's and 0's representing an indicator of treatment assignment. The causal treatment effect estimand is then obtained from the difference between the estimate of these two parameters. Here, we define the *Causal Average Treatment Effect* (CATE) as the estimated treatment effects in the first stratum (S_1):

$$CATE = \Delta_{CATE} = \hat{\Delta}_1 = \hat{\theta}_{11} - \hat{\theta}_{01} \tag{4.9}$$

The subscript in Δ (i.e. 1) represents stratum 1. The first and second subscript in θ correspond to treatment assignment and stratum respectively. We also define a *Weighted Average Treatment Effect* (WATE) as the overall average effect of the treatment that is evaluated with consideration of the stratum specific proportions. For this estimand, we only consider using strata with estimable quantities due to the missingness on ICE status.

$$WATE = \Delta_{WATE} = \hat{\Delta}_W = \left(\frac{\hat{\pi}_{11}}{\hat{\pi}_{11} + \hat{\pi}_{13}} \hat{\theta}_{11} + \frac{\hat{\pi}_{13}}{\hat{\pi}_{11} + \hat{\pi}_{13}} \hat{\theta}_{13} \right) - \left(\frac{\hat{\pi}_{01}}{\hat{\pi}_{01} + \hat{\pi}_{02}} \hat{\theta}_{01} + \frac{\hat{\pi}_{02}}{\hat{\pi}_{01} + \hat{\pi}_{02}} \hat{\theta}_{02} \right) \tag{4.10}$$

where the two subscripts in π represent the treatment assignment and stratum number respectively. Another standard treatment effect estimand that was calculated for comparison with CATE and WATE is the Intent-To-Treat (ITT) effect.

$$ITT = \Delta_{ITT} = \hat{\theta}_1 - \hat{\theta}_0 \tag{4.11}$$

In the context of ITT strategy, θ_1 and θ_0 represents the effects in the drug and control arm respectively regardless of intercurrent events. Using simulations, subject stratum membership, and hence strata proportions, are known. Similarly, the causal treatment effects values as well as their weighted causal treatment effects are also known.

$$\begin{aligned} \Delta_c &= \theta_{11} - \theta_{01} \\ \Delta_w &= \pi_1 \times \Delta_1 + \pi_2 \times \Delta_2 + \pi_3 \times \Delta_3 + \pi_4 \times \Delta_4 \end{aligned} \tag{4.12}$$

4.2.4 Performance Assessment

To assess the performance of the models, we compare the posterior mean square error (MSE) of the estimators. The posterior MSE in this case measures the average squared deviation of the estimated value from the true value and is useful in evaluating estimator efficiency. MSE is defined by

$$MSE_{\{Estimator\}} = E_{\Delta}\left[\left(\hat{\Delta} - \Delta\right)^2\right] \tag{4.13}$$

Here, the expectation is taken as the average across the number of iterations. From the simulated data, both the causal effects in the first strata (Δ_c) and the weighted treatment effect (Δ_w) are known. The MSE's for ITT, CATE and WATE are then calculated using the following functions

$$\begin{aligned} MSE\text{-}C_{ITT} &= E_{\Delta_c}\left[(\Delta_{ITT} - \Delta_c)^2\right] \\ MSE\text{-}W_{ITT} &= E_{\Delta_w}\left[(\Delta_{ITT} - \Delta_w)^2\right] \end{aligned} \tag{4.14}$$

For the principal stratification approach,

$$\begin{aligned} MSE_{CATE} &= E_{\Delta_c}\left[(\Delta_{CATE} - \Delta_c)^2\right] \\ MSE_{WATE} &= E_{\Delta_w}\left[(\Delta_{WATE} - \Delta_w)^2\right] \end{aligned} \tag{4.15}$$

A variant of MSE, Root Mean Square Error (RMSE) of the estimands, is used to visualize the performance of the estimands.

4.2.5 Motivating Example and Simulation Setup

Data for analysis was simulated based on partial descriptive statistics from a randomized control trial (Pollom et al., 2019), with various treatment effects of specific sizes imposed to aid in model

assessment. The distributions for simulating the data are provided in the appendix and are briefly defined below.

The primary outcome measure in the clinical trial (ELEMENT 5) was the change from baseline to 24 weeks in Hemoglobin A1c (HbA1c) for adults with type 2 diabetes. This was a prospective RCT to evaluate the safety and efficacy of the study drug known as LY2963016. Of clinical importance is reduction in HbA1c. All the data used in this study are based on simulation and should not be interpreted as results from the trial data. We simulated baseline characteristics in accordance with the distribution of the descriptive statistics from the study with some changes applied to other characteristics (Appendix: *Table A.1*). Intercurrent event status were then generated using arm-specific logistic function on baseline characteristics. This was followed by stratum allocation using the generated intercurrent event status and utilizing the alternative arm-specific logistic function. The alternative arm-specific logistic function generates counterfactual ICE probabilities. The parameters in the two logistic functions were chosen such that an approximate predetermined proportion for the four strata was achieved, which we also wanted to determine whether our modeling approach could recover. The responses (HbA1c) at week 24 were then generated using a regression function on baseline HbA1c and the specified values of treatment effects in each arm. Finally, subjects with intercurrent events at the week 24 were assigned missing values (assuming the intercurrent event precludes observation of HbA1c at week 24). Change in outcome was then calculated as the difference of response from baseline.

For performance assessment of the estimands, different simulation scenarios on the outcome (change in HbA1c) were considered in generating the final responses at week 24. We note that there are a multitude of possible scenarios and we only consider 5 possible scenarios that could affect treatment effect estimation and inference.

i. Homogeneous treatment effect across the four strata

ii. Heterogeneous treatment effect (a) (underestimated under ITT)

iii. Heterogeneous treatment effect (b) (overestimated under ITT)

iv. Reversed treatment effect

v. No treatment effect

For the *scenario i*, this assumes that the treatment effect is similar across the four strata (*Figure 4.2*). Here, it is assumed that if all potential outcomes were to be observed, the treatment effect would have been the same across the 4 strata. *scenario ii* and *iii* implies different effects of treatment depending on the subject's strata membership. This implies treatment effects are dependent on the subject's strata. The same applies to *scenario iv*, except that it considers other strata as having a reversed treatment effect. That is, subjects from one stratum will have positive treatment effects but subjects in other strata have negative effects from the treatment across the four strata. The last case is a scenario where there are no effects of the drug across the four strata. *Table 4.2* shows all the prespecified values for simulating the five scenarios.

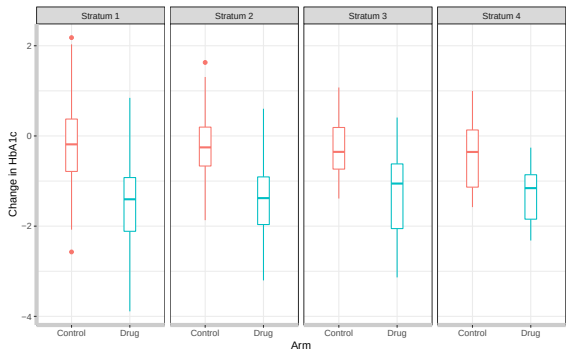
Table 4.2: Treatment effects simulation values

Effect Scenario	**STRATA**			
	Stratum 1 $\Delta_1(\theta_{01}, \theta_{11})$	Stratum 2 $\Delta_2(\theta_{02}, \theta_{12})$	Stratum 3 $\Delta_3(\theta_{03}, \theta_{13})$	Stratum 4 $\Delta_4(\theta_{04}, \theta_{14})$
i. Homogeneous	-1.25 (-0.25, -1.5)	-1.25 (-0.25, -1.5)	-1.25 (-0.25, -1.5)	-1.25 (-0.25, -1.5)
ii. Heterogeneous (a)	-1.25 (-0.25, -1.5)	-2.90 (-0.10, -3.0)	-3.75 (-0.25, -4.0)	-0.5 (-0.25, -0.75)
iii. Heterogeneous (b)	-1.25 (-0.25, -1.5)	-0.25 (-0.50, -0.75)	-0.75 (-0.25, -1.0)	-0.05 (-0.25, -0.3)
iv. Reversed	-1.25 (-0.25, -1.5)	-2.50 (1, -1.5)	3.0 (-1.5, 1.5)	0 (1.5, 1.5)
v. No Effect	0 (-0.25, -0.25)	0 (-0.25, -0.25)	0 (-0.25, -0.25)	0 (-0.25, -0.25)

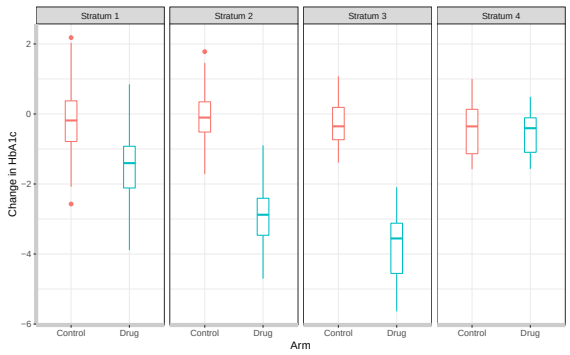
The values are treatment effects ($\Delta_S = \theta_{1S} - \theta_{0S}$); and the changes/effects in control and drug in parenthesis respectively (θ_{0S} - control effect stratum S; θ_{1S} - Drug effect stratum S)

We simulated data for the scenarios in *Table 4.2* using different sample sizes; 50, 100, 200, 400, and 1000. *Figure 4.2* provides a supplementary visualization of the scenarios using a sample of size 1000 and without imposing missingness in the subjects who had intercurrent events.

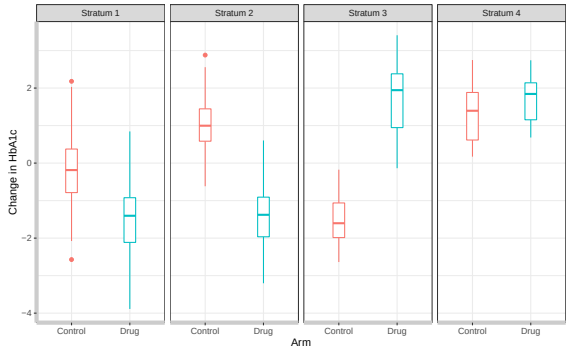
In the homogeneous group (*scenario i*), note the similarity of trend of boxplots across the four strata. This implies the treatment effects are the same across the four strata. As for *scenario ii*, there exist differences in treatment effects across the four strata with pronounced treatment effects



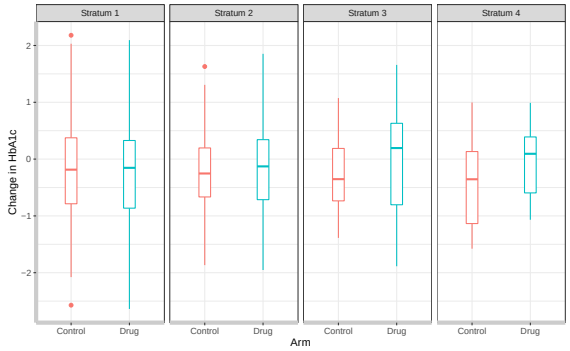
Scenario i. Homogeneous treatment effect



Scenario ii. Heterogeneous treatment effect (b)



Scenario iv. Reversed treatment effect



Scenario v. No treatment effect

Figure 4.2: Graphical illustration of simulation scenarios

in stratum 2 & 3 as compared to stratum 1 & 4. Visualization for *scenario iii*, (not shown), is similar to *scenario ii*, except with a different type of heterogeneity in treatment effects. As for the reversed case, we observe an opposite treatment effects for strata 1 & 2 compared to strata 3 & 4. Finally, the no effect *scenario v* exhibits cases where responses in controls are similar to responses in the drug across all four strata.

4.2.6 Bayesian Modeling Framework

We adopt MCMC techniques to estimate all parameters in the model using JAGS version 4.3.0 via `rjags package` in R version 3.6.0 (Plummer, 2003, 2019; R Core Team, 2019). The outcome (change in HbA1c) are modeled using a continuous distribution. HbA1c is regressed on the treatment indicator variables and the baseline HbA1c. In addition, we included baseline covariates in the model even though, only the baseline HbA1c was used to generate the final HbA1c. Intuitively,

this is an ANCOVA type model. Our approach entails two main parts, the latent strata and the response models. The latent strata principal stratification model is defined as follows

$$\begin{aligned} D_{01_i} &\sim Bernoulli(\pi_{01_i}) \\ D_{11_i} &\sim Bernoulli(\pi_{11_i}) \end{aligned} \tag{4.16}$$

where i is the subject index, D_{01} and D_{11} the observed intercurrent event status in control and drug respectively, with π_{01} and π_{11} their respective probabilities of experiencing an intercurrent event. The two equations are arm specific and of interest is estimating the attributes π_{01} and π_{11} for each subject. These estimates are obtained from the standard logistic function using two independent logit link function on baseline covariates

$$\begin{aligned} logit(\pi_{01_i}) &= \delta_{00} + \delta_{01}x_{01_i} + \cdots + \delta_{0q}x_{0q_i} \\ logit(\pi_{11_i}) &= \delta_{10} + \delta_{11}x_{11_i} + \ldots + \delta_{1q}x_{1q_i} \end{aligned} \tag{4.17}$$

As discussed previously, the two functions correspond to modeling intercurrent events for each arm independently, hence the reason for the two-subscript index on the δ parameters as well as the covariates. These result in treatment specific δ parameters. Covariates used in the model to simulate the events includes baseline HbA1c (X_1), baseline age (X_2), disease duration (X_3), and concomitant drug use (X_4). We assume these factors can influence a subject's ICE. During the data generation process, δ_{00} and δ_{10} control the proportions of generated ICE events in the control and drug arms respectively. As previously discussed, prediction of intercurrent events under the alternative assignment are calculated using the alternative treatment function and the model fit parameters. Therefore, all subjects have probabilities of experiencing an ICE under the alternative assignment. The latent strata membership for each subject i given their true assignment and observed ICE status is then modeled using a categorical distribution

$$S_i \sim Categorical(\pi_1, \pi_2, \pi_3, \pi_4) \tag{4.18}$$

where π's are based on the counterfactual predictive posterior distribution of experiencing an ICE using $\hat{\pi}_{01}$ and $\hat{\pi}_{11}$. During each iteration, some elements of S will be zero due to the fact that given a subject's true assignment and observed ICE status, they are considered to come from a mixture of only two strata.

After subject stratum allocation, the outcomes in each stratum are then modeled using a normal distribution

$$\begin{aligned} Y &\sim Normal\left(\mu_s, \sigma^2\right) \\ \mu_s &= \boldsymbol{Z}\theta_s + \boldsymbol{X}\beta \end{aligned} \tag{4.19}$$

where μ_s represent the estimated stratum specific means parameterized as a linear function of covariates. $\boldsymbol{Z}$ is an $n \times 2$ matrix containing indicator elements of treatment assignment; and $\boldsymbol{X}$ represent an $n \times q$ matrix of covariates to be adjusted for in the model. σ^2 is the associated variance of the outcome. In all models, diffuse normal priors for the latent strata model parameters δ as well as the regression parameters θ and β are defined. A noninformative inverse-gamma distribution is adopted for the variance parameter.

$$\begin{aligned} \delta &\sim Normal\,(0, 100) \\ \beta &\sim Normal\,(0, 100) \\ \sigma^2 &\sim Inverse\text{-}Gamma(0.001, 0.001) \end{aligned} \tag{4.20}$$

For the standard ITT estimation of treatment effects, the model and prior definitions follow the above specifications except that the intercurrent event status variable and hence the strata are ignored, otherwise all parameters are estimated in the same way for effective comparison. We recognize that the Bayesian framework treats missing outcomes in the ITT model as unknown quantities and are simulated during each iteration using MCMC.

4.3 Results

The posterior mean estimates as well as the 95% credible intervals for the treatment effect estimands (CATE, WATE, ITT) and their performance measures (MSEs) for *scenario i* are summa-

rized in *Table 4.3*. Ideally, the desired estimated values are the causal treatment effects among a homogeneous group of subjects in *stratum 1* (see *Table 4.1*) and the weighted average treatment effects across the four strata. The latter would be interpreted as reflecting the average treatment effects in the general population. In addition, treatment effects using the ITT approach is presented. Since a single value can only be estimated using ITT approach, we use this value to obtain both the $MSE\text{-}C_{ITT}$ associated with true true causal effect as well as $MSE\text{-}W_{ITT}$ associated with the true weighted treatment effects. $MSE\text{-}C_{ITT}$ is useful in assessing the magnitude of deviation between the estimated ITT and true causal effect among a homogeneous group of subjects. Similarly, $MSE\text{-}W_{ITT}$ reflects the level of deviation between the same ITT estimate with the true weighted treatment effects.

For *scenario i*, we assume the same effects of treatment across the four strata, (a difference of -1.25). From the simulated data, it is apparent that for small sample sizes, the estimate of the mean treatment effect is considerably farther than the true treatment effect. This is precisely captured using MSE, where we observe a consistent decrease with increasing sample size for all the estimands (*Figure 4.3*). We also observe comparable MSEs at different levels of sample sizes for all the treatment estimands. However, there is large variation that is associated with MSE for the CATE estimand as compared to WATE and ITT. It is important to note here that $MSE\text{-}C_{ITT} = MSE\text{-}W_{ITT}$ because a single value of ITT is estimated. In addition, this scenario assumes the same treatment effects across the four strata, hence, the $MSE\text{-}C_{ITT}$ and $MSE\text{-}W_{ITT}$ obtained in this scenario is denoted as MSE_{ITT} in the table

In the case of heterogeneity in treatment effects across the four strata, *scenario ii*, a substantial difference of estimates of treatment effects is observed for the different treatment effect estimands (*Table 4.4*). Firstly, we compare MSEs of CATE and ITT estimators. It is clear here that CATE has a consistently lower MSE (MSE_{CATE}) as compared to ITT ($MSE\text{-}C_{ITT}$) for all the sample sizes. On the other hand, the MSE for the WATE estimate (MSE_{WATE}) is slightly lower to some degree as compared to $MSE\text{-}W_{ITT}$. Secondly, from this scenario, we see that the estimates from WATE and ITT estimands will tend to overestimate the causal effects of the treatment as compared to the

Table 4.3: *Scenario i.* Homogeneous treatment effect ($\Delta_s = -1.25$); Weighted $\Delta = -1.25 (n = 1000)$

Sample Size	Estimator	Mean	SE	2.5%	50%	97.5%
50	CATE	-1.616	0.248	-2.114	-1.614	-1.135
	WATE	-1.588	0.239	-2.058	-1.588	-1.117
	ITT	-1.597	0.238	-2.062	-1.599	-1.123
	MSE_{CATE}	0.196	0.208	0.001	0.133	0.747
	MSE_{WATE}	0.171	0.183	< 0.001	0.115	0.654
	MSE_{ITT}	0.177	0.184	< 0.001	0.123	0.66
100	CATE	-1.425	0.206	-1.832	-1.425	-1.022
	WATE	-1.498	0.191	-1.874	-1.497	-1.123
	ITT	-1.481	0.192	-1.86	-1.48	-1.102
	MSE_{CATE}	0.073	0.095	< 0.001	0.038	0.34
	MSE_{WATE}	0.098	0.108	< 0.001	0.062	0.39
	MSE_{ITT}	0.09	0.103	< 0.001	0.055	0.372
200	CATE	-1.315	0.179	-1.652	-1.32	-0.951
	WATE	-1.252	0.129	-1.504	-1.251	-0.998
	ITT	-1.255	0.131	-1.513	-1.255	-0.998
	MSE_{CATE}	0.036	0.05	< 0.001	0.017	0.176
	MSE_{WATE}	0.017	0.024	< 0.001	0.007	0.085
	MSE_{ITT}	0.017	0.025	< 0.001	0.008	0.087
400	CATE	-1.307	0.126	-1.548	-1.308	-1.058
	WATE	-1.225	0.085	-1.392	-1.226	-1.059
	ITT	-1.224	0.088	-1.395	-1.223	-1.053
	MSE_{CATE}	0.019	0.026	< 0.001	0.009	0.092
	MSE_{WATE}	0.008	0.011	< 0.001	0.004	0.039
	MSE_{ITT}	0.008	0.012	< 0.001	0.004	0.042
1000	CATE	-1.246	0.08	-1.397	-1.247	-1.086
	WATE	-1.272	0.054	-1.38	-1.272	-1.165
	ITT	-1.271	0.056	-1.381	-1.271	-1.163
	MSE_{CATE}	0.006	0.009	< 0.001	0.003	0.032
	MSE_{WATE}	0.003	0.005	< 0.001	0.002	0.018
	MSE_{ITT}	0.004	0.005	< 0.001	0.002	0.018

CATE. Both WATE and ITT exhibit an almost identical estimates of treatment effects, except that WATE estimate shows lower variability than the ITT estimate.

For the other three scenarios, we provide a similar, structured tabular results in the appendix. In the case of heterogeneous treatment effects in *scenario iii*, we observe the same trend in performance that is comparable to *scenario ii*. However, both WATE and ITT estimates of treatment effect will tend to underestimate the causal treatment effect for this scenario using moderate to large sample size (*Table B.1* - Appendix). With reversed treatment effects across the four strata

Table 4.4: *Scenario ii.* Heterogeneous treatment effect (a) ($\Delta_1 = -1.25,\ \Delta_2 = -2.90,\ \Delta_3 = -3.75,\ \Delta_4 = -0.50$); Weighted $\Delta = -1.654$ ($n = 1000$)

Sample Size	Estimator	Mean	SE	2.5%	50%	97.5%
50	CATE	-2.266	0.351	-2.953	-2.266	-1.574
	WATE	-2.27	0.348	-2.953	-2.269	-1.583
	ITT	-2.271	0.346	-2.947	-2.274	-1.583
	MSE_{CATE}	1.156	0.735	0.106	1.032	2.899
	$MSE\text{-}C_{ITT}$	1.162	0.724	0.112	1.048	2.881
	MSE_{WATE}	0.393	0.405	0.001	0.273	1.451
	$MSE\text{-}W_{ITT}$	0.393	0.4	0.001	0.278	1.439
100	CATE	-1.528	0.211	-1.944	-1.528	-1.112
	WATE	-1.63	0.196	-2.016	-1.63	-1.242
	ITT	-1.607	0.198	-1.998	-1.606	-1.215
	MSE_{CATE}	0.122	0.134	< 0.001	0.079	0.481
	$MSE\text{-}C_{ITT}$	0.166	0.152	0.001	0.127	0.56
	MSE_{WATE}	0.039	0.057	< 0.001	0.018	0.2
	$MSE\text{-}W_{ITT}$	0.039	0.057	< 0.001	0.018	0.204
200	CATE	-1.258	0.189	-1.617	-1.264	-0.875
	WATE	-1.675	0.131	-1.934	-1.675	-1.418
	ITT	-1.683	0.172	-2.022	-1.683	-1.346
	MSE_{CATE}	0.036	0.05	< 0.001	0.017	0.177
	$MSE\text{-}C_{ITT}$	0.217	0.155	0.01	0.188	0.597
	MSE_{WATE}	0.032	0.04	< 0.001	0.017	0.142
	$MSE\text{-}W_{ITT}$	0.042	0.057	< 0.001	0.02	0.204
400	CATE	-1.234	0.114	-1.456	-1.235	-1.009
	WATE	-1.372	0.087	-1.542	-1.372	-1.2
	ITT	-1.383	0.096	-1.57	-1.382	-1.196
	MSE_{CATE}	0.013	0.019	< 0.001	0.006	0.067
	$MSE\text{-}C_{ITT}$	0.027	0.029	< 0.001	0.018	0.102
	MSE_{WATE}	0.097	0.053	0.017	0.089	0.222
	$MSE\text{-}W_{ITT}$	0.092	0.057	0.01	0.084	0.226
1000	CATE	-1.291	0.08	-1.45	-1.29	-1.137
	WATE	-1.505	0.055	-1.612	-1.505	-1.398
	ITT	-1.503	0.063	-1.627	-1.503	-1.381
	MSE_{CATE}	0.008	0.011	< 0.001	0.004	0.04
	$MSE\text{-}C_{ITT}$	0.068	0.032	0.017	0.064	0.142
	MSE_{WATE}	0.025	0.017	0.002	0.022	0.066
	$MSE\text{-}W_{ITT}$	0.027	0.02	0.001	0.023	0.075

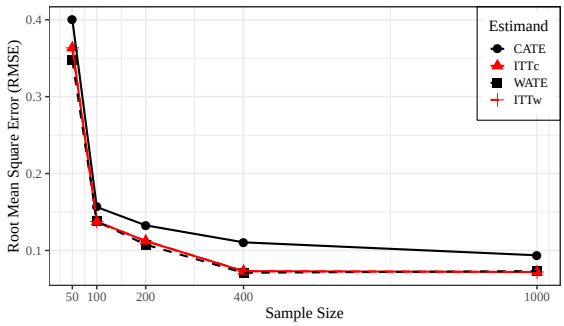
(*scenario iv*) the results from our simulation do not exhibit much deviation when comparing ITT to CATE as well as ITT to WATE (*Table B.2* - Appendix). Similarly, when there are no treatment effects across the four strata (*scenario v*) we observe the same trend in terms of estimates as well as performance (*Table B.3* - Appendix). This is comparable to the homogeneity in treatment ef-

fect as in *scenario i*. We also present the plots of RMSE for all scenarios (*Figure 4.3*). RMSE is calculated as the square root of MSE. There are two factors evaluated by the models: the effect within the homogenous group of subjects, and the weighted effects from all strata. In all the plots, CATE is compared to ITTc (from associated causal ITT difference) while WATE is compared to ITTw (from associated weighted ITT difference). The line type in *Figure 4.3* differentiates the pair of estimands to compare. In all cases, we note the substantial decrease in RMSE with increasing sample sizes. ITTc (solid red) performs better in both *scenario i* and *v* as compared to CATE (solid black). With increasing sample size, the pattern is consistently lower using the ITT estimand. This is similarly observed by comparing ITTw (dashed red) to WATE (dashed black). In these two cases, ITT models benefits from the fact that it is estimating the same parameter with a larger sample size as compared to the other two estimands. In *scenario ii*, both CATE and WATE exhibit better performance as compared to estimates obtained from the ITT especially with increasing sample sizes. In case of *scenario iii*, we observe large values of RMSE for the ITT model compared to WATE and this trend is also observed in *scenario iv* which appears unstable even with increasing sample size when the treatment effects are reversed.

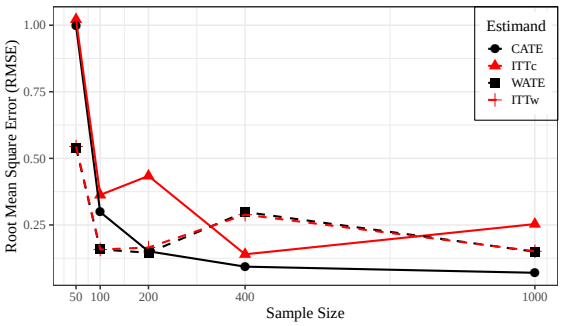
4.4 Discussion

In this study, we developed a framework of assessing treatment effects for clinical trials that are prone to intercurrent events. The goal of randomized clinical trial is to obtain the causal effect of the treatment through randomization hence controlling for potential confounders. However, certain factors do occur in the course of a clinical trial that precludes observing the outcome of interest during trial follow up. Using MSE, we have shown that, utilization of principal stratification with strata predictive covariates to estimate treatment effects demonstrates a better performance than standard approaches that adopt the ITT strategy specifically when there is presence of heterogeneity in treatment effects.

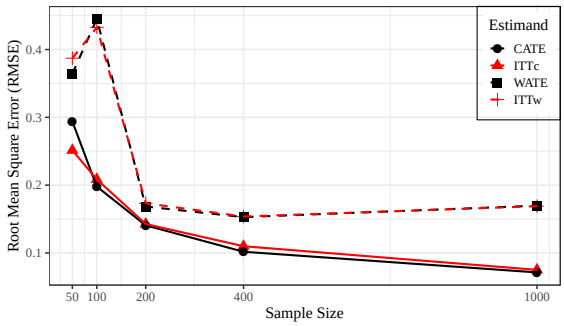
Two simulation scenarios: *scenario ii* & *iii*, provide a clear illustration on how causal treatment effects can be overestimated or underestimated when there are intercurrent events and the standard



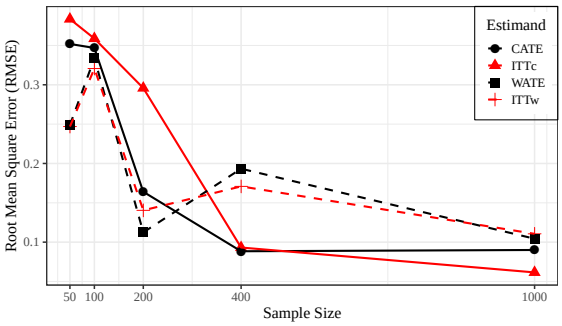
Scenario i. Homogeneous treatment effect



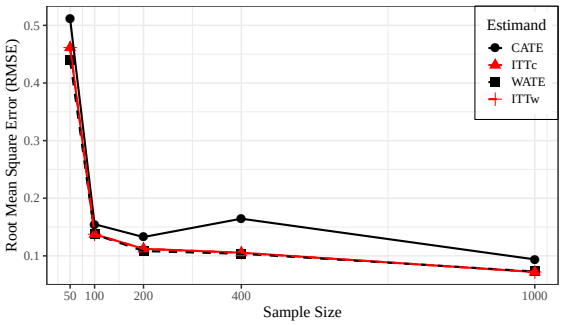
Scenario ii. Heterogeneous treatment effect (a)



Scenario iii. Heterogeneous treatment effect (b)



Scenario iv. Reversed treatment effect



Scenario v. No treatment effect

Figure 4.3: **RMSE plots for estimator comparison:** Plots of Root Mean Square Error (RMSE) for all the scenarios. The lines with same color correspond to the same-type model i.e. ITT or PS. Comparisons should be performed on similar line type i.e. CATE-Causal Average Treatment Effect with ITTc-Intent-to-treat causal comparison (Solid black vs solid red): WATE- Weighted Average Treatment Effect from principal Stratification with ITTw-Intent-to-treat weighted effects comparison (dashed black vs dashed red). NOTE: Horizontal axis not drawn to scale.

ITT strategy is adopted. For instance, in the heterogeneous treatment effect case we note that the ITT will overestimate the causal effect of the treatment *scenario ii*, while the principal stratification

framework provides a robust way of assessing the treatment effects using two estimands, CATE and WATE. The CATE yields a lower estimate of treatment effect than ITT that does not overestimate the treatment effect. In addition, CATE has a lower RMSE (*Figure 4.3*). A similar case but with underestimated treatment effects when using ITT is observed in *scenario iii.* This will cause approval of drugs that the risk outweighs their benefits leading to unnecessary harm to more patients that might not benefit from the treatment. On the other hand, underestimated treatment effects may result in early clinical trial termination leading to patients who would have benefitted from such drugs being denied of potentially life-saving interventions. Both scenarios warrant in-depth assessment and this framework provides a complementary means to assess effects of the treatments that might inform critical decisions. For instance, an overestimated treatment effect is considerably important for regulatory agencies when making drug approval decisions. In addition, it could be extremely important for investigators to assess the effectiveness of treatments in subgroup of patients during late-phase clinical trials or post-marketing of the new drug. On the other hand, an underestimated treatment effect could be crucial in drug development process for investigators especially if the threshold of efficacy is not achieved using standard ITT methods.

Compared to the standard ITT strategy, we obtained more precise estimates using a weighted treatment effects approach in almost all the scenarios. A feature that is lacking using the standard ITT is the identification of the causal treatment effects if there is heterogeneity in treatment effects across subjects. Our approach provides a way to establish both the causal effects based on a predicted homogeneous group of subjects in addition to weighted treatment effects based on the entire pool of subjects under study. We note that, in the case of the weighted average treatment effects (WATE), the results can only be obtained using observable strata. That is, if ICE results in missing outcomes, then effects in stratum 4 will be highly influenced by the prior definition. Non-informative priors will most likely have undesired effects on this stratum. Therefore, calculation of the WATE in our approach involved using only those strata with estimable quantities. On the other hand, cases of intercurrent events that result in non-missing outcomes (e.g. treatment switch or use of rescue medication) do not affect calculation of the WATE since all quantities will be estimable

from all strata. What is interesting from our simulation examples is the fact that MSE using a weighted approach is less than the MSE of the ITT strategy in almost all scenarios (*Figure 4.3*). This suggest that using predictive covariates for strata prediction post-randomization is beneficial in estimation of treatment effects. When the treatment effects are similar across the four strata (*scenario i* and *scenario v*, see *Table 4.3* & *Table B.3*), we observed larger variation in MSE for CATE estimand as compared to WATE and ITT. This is attributed to the fact that CATE is estimated with a smaller sample size compared to WATE and ITT.

In designing a new clinical trial based on historical data or real-world-data (RWD) from similar trials, the latent strata membership model can be useful in defining inclusion/exclusion criteria where subjects who might potentially experience an intercurrent event are flagged pre-randomization. This is similarly applicable during screening of participants for the succeeding phases of a clinical trial. Specifically, it ensures the treatment effect is estimated using a homogeneous group of subjects. During interim analysis for flexible designs like adaptive clinical trial designs, this framework can prove to be beneficial in facilitating trial modification through efficacy evaluation and identification of patient population who have the potential of reaping the greatest benefits from the drugs whilst limiting exposure to harm in patient populations who would not benefit.

The factors included in the latent strata models do not need to have a significant marginal association with the intercurrent event variable in order to be included in the latent strata models since their parameter estimates will be shrunk to 0 during MCMC iterations. In addition, for the two latent strata models, the predictive covariates used to define the two models need not be identical. That is, one can have different number of predictor covariates in the latent strata for control arm compared to the treatment arm. Similarly, the variables used to predict strata can be different from the covariates used in the modeling the actual outcome.

In this study, we do not impose any assumptions of monotonicity or the exclusion restrictions. We recognize that while these assumptions are regularly considered when using the principal stratification framework (Imbens & Rubin, 1997; Page et al., 2015), they are sometimes hard to test. In practice, some assumptions regarding monotonicity can be helpful in mitigating issues of iden-

tifiability (Magnusson et al., 2019). In all simulations, evaluation of performance on estimators involved comparing the MSE of the estimands between the principal stratification approach and the standard ITT strategy without imposing any assumptions.

Bayesian analysis require prior definition on parameters. Here, we used non-informative priors, however, if evidence exists from prior studies or RWD, precision on treatment effects estimates could benefit from informative prior definition. In our analysis we utilized an overall variance parameter across the four strata. Other cases might consider different variances for each stratum but this will result in convergence issues especially when using a vague prior for the variance parameters in each stratum. In addition, this may result in label switching issues (Malsiner-Walli et al., 2017). In our model definition, the ANCOVA model is adopted in modeling the outcome (change in HbA1c) with adjustment for baseline measure (Clifton & Clifton, 2019; Liu et al., 2009). Given that this is an RCT, the Bayesian approach will shrink the covariate mean effects to zero if they do not contribute to the response.

Our approach relies on an aggregate of intercurrent events, or in other terms, single type intercurrent event. We recognize this limitation given that there is a finite diverse type of intercurrent events. Although this limitation is not addressable with our current approach, future work will involve extension of our approach to handle multiple types of intercurrent events simultaneously in assessing causal treatment effects. In addition, we only considered cases where ICE results in unobserved outcomes. Other cases may involve observed outcomes even after an intercurrent event, for example, when subjects use rescue medication or switched treatments. While this approach can still be used, further assessments needs to be performed to assess the impact on the estimate of causal treatment effect.

4.5 Conclusion

This approach provides additional benefits of inference beyond the ITT effect since we can get a better estimate of the true magnitude of treatment effect in addition to the standard ITT effect that can improve the inference on estimated treatment effect as well as identifying subjects or subgroup

of population who will benefit from the drug. It also creates a framework to flag subjects who would not benefit from the drug especially for adaptive clinical trials hence, limiting harm that may be caused by the treatment. Study drugs that do not show efficacy in trials may be due to conduct of the trials on non-responders, however, there may be a group of patients that would benefit from the drug. Therefore, screening for inclusion in the trial can be informed by such an approach on historical data or RWD investigating similar compounds. This could prove to be of great benefit in the design and analysis of clinical trials as well as the inference and interpretation of treatment effect estimates. Although this study is based on simulated data on RCT, it can also be applied to observational data. Finally, it would be useful to know which subjects would benefit most from treatments, and our approach could help provide helpful information in regards to drug efficacy on a subgroup of patients, leading to targeted interventions.

Chapter 5

Summary and Future Directions

This dissertation work provides a framework for enhanced understanding and evaluation of clinical outcomes. First, we extended application of a machine learning algorithm (Bayesian Networks) to model time-to-event data, and second, proposed a new approach for estimating treatment effects in clinical trials with intercurrent events. These approaches provide a path to understanding and improving inference in clinical outcomes especially with the evolution in the diversity of data types collected from patients and the need to gain a better understanding of the structural relationship between variables over time as well as obtaining unbiased estimates of treatment effects in randomized clinical trials with intercurrent events.

In Chapter 2, we introduced the new approach of analyzing time-to-event data using Bayesian Networks, which allows for flexibility in structural variable relationships at different periods. The data is first pre-processed by discretizing survival times into binary outcomes which are then modeled sequentially using Bayesian Networks. This analysis approach is primarily driven by the increasing diversity in data types and the need to understand variable relationships at different time points. The new analysis framework, survival Bayesian Networks, demonstrated a better or comparable performance as compared to logistic regression for clinical characteristics and the integrated clinical characteristics and gene expression data. However, this was not the case for all cancer types where logistic regression exhibited better performance than the survival Bayesian Networks for integrated data in kidney cancer. This poor performance on kidney cancer was attributed to overfitting with the inclusion of the selected genes. We provided a sample of the learned networks for the first time period using integrated data. The flexibility in learning different structural relationships between variables at different time periods might potentially inform patient care and the development of targeted therapies. Our method further mitigates challenges associated with proportional hazard assumptions using the standard Cox proportional hazards model. Currently, this approach requires pre-specification of periodic interval cutoffs for creating periodic binary out-

comes. However, future research will consider other strategies to identify the time-points to adopt as interval cutoffs. Also, approaches to prune the network to improve efficiency and accuracy of modeling and avoid cases of overfitting will be utilized. Another future consideration involves an extension to model competing risk events (e.g. relapse, remission, death, e.t.c).

In Chapter 3, we provided a review of applications of predictive modeling and their benefits to clinical trials. Randomized clinical trials are associated with high cost and long durations with high proportions of failures. In this chapter, we highlighted some aspects of clinical trials that have benefited or have the potential to benefit from applicability of predictive modeling. Identified areas encompass stages of clinical trial process from trial design, patient recruitment, trial optimization to safety and efficacy evaluation.

In Chapter 4, we expanded the principal stratification framework (Frangakis & Rubin, 2002) to estimate treatment effects in clinical trials having post-randomization/intercurrent events. Clinical trials are prone to post-randomization events that can have an impact on evaluation and interpretation of the estimated treatment effects. Intercurrent events e.g. treatment discontinuation or death, leads to missing outcomes at the end of the trial while other intercurrent events e.g. treatment switch or use of rescue medication, will have outcomes observed at the end of the trial. Our analyses are based on simulations and we considered cases of missing outcomes after an intercurrent event. In this chapter, two estimators are presented; Causal Average Treatment Effect (CATE) and Weighted Average Treatment Effect (WATE). CATE ensures estimation of treatment effects using a homogeneous group of subjects. WATE, on the other hand, represents an estimation of overall treatment effects similar to the ITT approach but with the effects adjusted for by estimated strata specific proportions. These two estimators provide a robust means of estimating treatment effects or can be used to complement analysis based on the Intent-to-treat (ITT) strategy. For both CATE and WATE, subjects are stratified based on their observed baseline covariate values, treatment assignment, observed intercurrent event and their propensity of experiencing an intercurrent event under the alternative treatment assignment. Using simulations, both CATE and WATE demonstrated lowest estimator MSEs, specifically with heterogeneity in treatment effects across

the strata. When the treatment effect is homogenous across strata, using ITT strategy was better than using CATE. However, WATE and ITT did not exhibit a difference in performance. Moreover, WATE had the lowest variability as compared to ITT. In this chapter, we considered a combination of multiple intercurrent events when defining the intercurrent event variable. Future studies will look into strategies to address the different types of intercurrent events when estimating the treatment effects.

References

Abdallah, K., Hugh-Jones, C., Norman, T., Friend, S., & Stolovitzky, G. (2015). The prostate cancer dream challenge: A community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer. *Oncologist*, 20(5), 459–60.

Abrahams, E. (2008). Right drug-right patient-right time: personalized medicine coalition. *Clin Transl Sci*, 1(1), 11–2.

Aroda, V. R., Saugstrup, T., Buse, J. B., Donsmark, M., Zacho, J., & Davies, M. J. (2019). Incorporating and interpreting regulatory guidance on estimands in diabetes clinical trials: The pioneer 1 randomized clinical trial as an example. *Diabetes Obes Metab*, 21(10), 2203–2210.

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., Konig, F., & Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PLoS One*, 13(10), e0205971.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E., & O'Connor, P. J. (2015). Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29(4), 1033–1069.

Barnard, K. D., Dent, L., & Cook, A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC Med Res Methodol*, 10, 63.

Bate, A., Juniper, J., Lawton, A. M., & Thwaites, R. M. (2016). Designing and incorporating a real world data approach to international drug development and use: what the uk offers. *Drug discovery today*, 21(3), 400–405.

Berry, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statist. Sci.*, 19(1), 175–187.

Berry, D. A. (2006). Bayesian clinical trials. *Nat Rev Drug Discov*, 5(1), 27–36.

Brady, J. J., Chuang, C. H., Greenside, P. G., Rogers, Z. N., Murray, C. W., Caswell, D. R., Hartmann, U., Connolly, A. J., Sweet-Cordero, E. A., Kundaje, A., & Winslow, M. M. (2016). An arntl2-driven secretome enables lung adenocarcinoma metastatic self-sufficiency. *Cancer Cell*, 29(5), 697–710.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199–231.

Broglio, K. R., Stivers, D. N., & Berry, D. A. (2014). Predicting clinical trial results based on announcements of interim analyses. *Trials*, 15, 73.

Chang, F., Steelman, L. S., Lee, J. T., Shelton, J. G., Navolanic, P. M., Blalock, W. L., Franklin, R. A., & McCubrey, J. A. (2003). Signal transduction mediated by the ras/raf/mek/erk pathway from cytokine receptors to transcription factors: potential targeting for therapeutic intervention. *Leukemia*, 17(7), 1263–93.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*, 3(3), 243–50.

Cheng, F. & Zhao, Z. (2014). Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc*, 21(e2), e278–86.

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4), e1006076.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3, 140–140.

Clifton, L. & Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials*, 20(1), 43.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC Med*, 13, 1.

Consortium, E. P. (2011). A user‘s guide to the encyclopedia of dna elements (encode). *PLoS Biol*, 9(4), e1001046.

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

Cui, Y., Murphy, B., Gentilcore, A., Sharma, Y., Minasian, L. M., Kramer, B. S., Coates, P. M., Gohagan, J. K., Klenk, J., & Tidor, B. (2014). Multilevel modeling and value of information in clinical trial decision support. *BMC Syst Biol*, 8, 6.

DiMasi, J. A., Hermann, J. C., Twyman, K., Kondru, R. K., Stergiopoulos, S., Getz, K. A., & Rackoff, W. (2015). A tool for predicting regulatory approval after phase ii testing of new oncology compounds. *Clin Pharmacol Ther*, 98(5), 506–13.

Ding, P. & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3), 757–777.

Elangovan, I. M., Vaz, M., Tamatam, C. R., Potteti, H. R., Reddy, N. M., & Reddy, S. P. (2018). Fosl1 promotes kras-induced lung cancer through amphiregulin and cell survival gene regulation. *American journal of respiratory cell and molecular biology*, 58(5), 625–635.

Erikainen, S. & Chan, S. (2019). Contested futures: envisioning “personalized,” “stratified,” and “precision” medicine. *New Genetics and Society*, 38(3), 308–330.

Escudero, J., Zajicek, J. P., Ifeachor, E., & Alzheimer’s Disease Neuroimaging, I. (2011). Machine learning classification of mri features of alzheimer’s disease and mild cognitive impairment

subjects to reduce the sample size in clinical trials. *Conf Proc IEEE Eng Med Biol Soc*, 2011, 7957–60.

Exarchos, T. P., Rigas, G., Goletsis, Y., Stefanou, K., Jacobs, S., Trivella, M. G., & Fotiadis, D. I. (2014). A dynamic bayesian network approach for time-specific survival probability prediction in patients after ventricular assist device implantation. *Conf Proc IEEE Eng Med Biol Soc*, 2014, 3172–5.

Ezzati, A. & Lipton, R. B. (2020). Machine learning predictive models can improve efficacy of clinical trials for alzheimer's disease. *Journal of Alzheimer's Disease*, 74, 55–63.

Feller, A., Mealli, F., & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, 42(6), 726–758.

Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary clinical trials communications*, 11, 156–164.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–9.

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *Am J Epidemiol*, 173(7), 761–7.

Gajewski, B. J., Simon, S. D., & Carlson, S. E. (2008). Predicting accrual in clinical trials with bayesian posterior predictive distributions. *Stat Med*, 27(13), 2328–40.

Garrison, L. P., Neumann, P. J., Erickson, P., Marshall, D., & Mullins, C. D. (2007). Using real-world data for coverage and payment decisions: The ispor real-world data task force report. *Value in Health*, 10(5), 326–335.

Goodman, M. S., Li, Y., & Tiwari, R. C. (2011). Detecting multiple change points in piecewise constant hazard functions. *J Appl Stat*, 38(11), 2523–2532.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*, 7, 496.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *N Engl J Med*, 375(12), 1109–12.

Guinney, J., Wang, T., Laajala, T. D., Winner, K. K., et al. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *The Lancet Oncology*, 18(1), 132–142.

Gullick, N. J., Mian, A. N., Ibrahim, F., Walker, D., Hassell, A., Kiely, P. D. W., Walsh, D. A., Young, A., Scott, D. L., & Investigators, T. P. (2017). Predicting responses in patients with rheumatoid arthritis to disease-modifying agents using baseline clinical data. *Clin Exp Rheumatol*, 35(5), 810–815.

Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(2), 188–195.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspect Clin Res*, 2(3), 109–12.

Gupta, S. K. (2012). Use of bayesian statistics in drug development: Advantages and challenges. *International journal of applied & basic medical research*, 2(1), 3–6.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., & Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics*, 2009.

Harrison, R. K. (2016). Phase ii and phase iii failures: 2013-2015. *Nat Rev Drug Discov*, 15(12), 817–818.

Hernandez, I. & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes. *American Journal of Health-System Pharmacy*, 74(18), 1494–1500.

Holzinger, E. R. & Ritchie, M. D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, 13(2), 213–22.

Huang, H., Chen, J., Ding, C. M., Jin, X., Jia, Z. M., & Peng, J. (2018). Lncrna nr2f1-as1 regulates hepatocellular carcinoma oxaliplatin resistance by targeting abcc1 via mir-363. *J Cell Mol Med*, 22(6), 3238–3245.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.

Imbens, G. W. & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.*, 25(1), 305–327.

International Council for Harmonization (2019). Addendum on estimands and sensitivity analysis in clinical trials: To the guidelines on statistical principles for clinical trials.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC medical research methodology*, 17(1), 162–162.

Jiang, Y., Simon, S., Mayo, M. S., & Gajewski, B. J. (2015). Modeling and validating bayesian accrual models on clinical data and simulations using adaptive priors. *Stat Med*, 34(4), 613–29.

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Medicinal Chemistry*, 3(6), 647–663.

Juengst, E. T. & McGowan, M. L. (2018). Why does the shift from "personalized medicine" to "precision health" and "wellness genomics" matter? *AMA J Ethics*, 20(9), E881–890.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.

Keene, O. N. (2019). Strategies for composite estimands in confirmatory clinical trials: Examples from trials in nasal polyps and steroid reduction. *Pharm Stat*, 18(1), 78–84.

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I. H., & Kim, H. J. (2019). Deep learning-based survival prediction of oral cancer patients. *Sci Rep*, 9(1), 6994.

Kopcke, F., Lubgan, D., Fietkau, R., Scholler, A., Nau, C., Sturzl, M., Croner, R., Prokosch, H. U., & Toddenroth, D. (2013). Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Med Inform Decis Mak*, 13, 134.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, 13, 8–17.

Krams, M., Lees, K. R., Hacke, W., Grieve, A. P., Orgogozo, J. M., Ford, G. A., & Investigators, A. S. (2003). Acute stroke therapy by inhibition of neutrophils (astin): an adaptive dose-response study of uk-279,276 in acute ischemic stroke. *Stroke*, 34(11), 2543–8.

Kueffner, R., Zach, N., Bronfeld, M., Norel, R., Atassi, N., Balagurusamy, V., & Others (2019). Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci Rep*, 9(1), 690.

Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1), 31–57.

Li, S., Zheng, K., Pei, Y., Wang, W., & Zhang, X. (2019). Long noncoding rna nr2f1-as1 enhances the malignant properties of osteosarcoma by increasing forkhead box a1 expression via sponging of microrna-483-3p. *Aging (Albany NY)*, 11(23), 11609–11623.

Li, Y. Y. & Jones, S. J. (2012). Drug repositioning for personalized medicine. *Genome Med*, 4(3), 27.

Lin, E. & Lane, H. Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomark Res*, 5, 2.

Liu, G. F., Lu, K., Mogg, R., Mallick, M., & Mehrotra, D. V. (2009). Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in medicine*, 28(20), 2509–2530.

Ma, B., Geng, Y., Meng, F., Yan, G., & Song, F. (2020). Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. *Journal of Cancer*, 11(5), 1288–1298.

Magnusson, B. P., Schmidli, H., Rouyrre, N., & Scharfstein, D. O. (2019). Bayesian inference for a principal stratum estimand to assess the treatment effect in a subgroup characterized by postrandomization event occurrence. *Stat Med*, 38(23), 4761–4771.

Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2017). Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2), 285–295.

Man, J., Zhang, X., Dong, H., Li, S., Yu, X., Meng, L., Gu, X., Yan, H., Cui, J., & Lai, Y. (2019). Screening and identification of key biomarkers in lung squamous cell carcinoma by bioinformatics analysis. *Oncol Lett*, 18(5), 5185–5196.

Mattei, A. & Mealli, F. (2007). Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63(2), 437–46.

Meier, R., Graw, S., Usset, J., Raghavan, R., Dai, J., Chalise, P., Ellis, S., Fridley, B., & Koestler, D. (2016). An ensemble-based cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mcrpc) patients. *F1000Res*, 5, 2677.

Menard, T., Barmaz, Y., Koneswarakantha, B., Bowling, R., & Popko, L. (2019). Enabling data-driven clinical quality assurance: Predicting adverse event reporting in clinical trials using machine learning. *Drug Saf*, 42(9), 1045–1053.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann Intern Med*, 162(1), W1–73.

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., & Greyson, D. (2011). The cost of drug development: a systematic review. *Health Policy*, 100(1), 4–17.

Mudaranthakam, D. P., Thompson, J., Hu, J., Pei, D., Chintala, S. R., Park, M., Fridley, B. L., Gajewski, B., Koestler, D. C., & Mayo, M. S. (2018). A curated cancer clinical outcomes database (c3od) for accelerating patient recruitment in cancer clinical trials. *JAMIA Open*, 1(2), 166–171.

Mulder, N. J., Akinola, R. O., Mazandu, G. K., & Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and Structural Biotechnology Journal*, 11(18), 1–10.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. *Journal of Cheminformatics*, 5(1), 30.

Neums, L., Meier, R., Koestler, D., & Thompson, J. (2020). Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25, 415.

Ni, Y., Kennebeck, S., Dexheimer, J. W., McAneney, C. M., Tang, H., Lingren, T., Li, Q., Zhai, H., & Solti, I. (2015). Automated clinical trial eligibility prescreening: increasing the efficiency of

patient identification for clinical trials in the emergency department. *J Am Med Inform Assoc*, 22(1), 166–78.

Obermeyer, Z. & Emanuel, E. J. (2016). Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*, 375(13), 1216–9.

Odondi, L. & McNamee, R. (2013). Applying optimal model selection in principal stratification for causal inference. *Stat Med*, 32(11), 1815–28.

Page, D., Costa, V. S., Natarajan, S., Barnard, A., Peissig, P., & Caldwell, M. (2012). Identifying adverse drug events by relational learning. *Proc Conf AAAI Artif Intell*, 2012, 790–793.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, 36(4), 514–531.

Pappalardo, F., Russo, G., Tshinanu, F. M., & Viceconti, M. (2019). In silico clinical trials: concepts and early adoptions. *Brief Bioinform*, 20(5), 1699–1708.

Passos, I. C. & Mwangi, B. (2018). Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. *Mol Psychiatry*.

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 21(6), 872–875.

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier Science.

Pencina, M. J. & D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*, 23(13), 2109–23.

Personalized Medicine Coalition (2014). *The Case for Personalized Medicine*. Report.

Persson, I. & Khamis, H. (2005). Bias of the cox model hazard ratio. *Journal of Modern Applied Statistical Methods*, 4(1), 10.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., & West, M. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A*, 101(22), 8431–6.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124 (pp. 1–10).: Vienna, Austria.

Plummer, M. (2019). rjags: Bayesian graphical models using mcmc. *R package version*, 4-10.

Pollom, R. K., Ilag, L. L., Lacaya, L. B., Morwick, T. M., & Ortiz Carrasquillo, R. (2019). Lilly insulin glargine versus lantus((r)) in insulin-naive and insulin-treated adults with type 2 diabetes: A randomized, controlled trial (element 5). *Diabetes Ther*, 10(1), 189–203.

Porcher, R., Jacot, J., Wunder, J. S., & Biau, D. J. (2019). Identifying treatment responders using counterfactual modeling and potential outcomes. *Statistical Methods in Medical Research*, 28(10-11), 3346–3362.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj, U., Aier, I., Semwal, R., & Varadwaj, P. K. (2017). Identification of novel dysregulated key genes in breast cancer through high throughput chip-seq data analysis. *Scientific reports*, 7(1), 3229–3229.

Ratitch, B., Bell, J., Mallinckrodt, C., Bartlett, J. W., Goel, N., Molenberghs, G., O'Kelly, M.,

Singh, P., & Lipkovich, I. (2020). Choosing estimands in clinical trials: Putting the ich e9(r1) into practice. *Ther Innov Regul Sci*, 54(2), 324–341.

RCT Duplicate (2020). Effectiveness research with real world data to support fda's regulatory decision making.

Renfro, L. A., Mallick, H., An, M. W., Sargent, D. J., & Mandrekar, S. J. (2016). Clinical trial designs incorporating predictive biomarkers. *Cancer Treat Rev*, 43, 74–82.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015a). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, 16(2), 85–97.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015b). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–40.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5), 656–666.

Roy, J., Hogan, J. W., & Marcus, B. H. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, 9(2), 277–89.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371), 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961–962.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 322–331.

Russell, S. J. & Norvig, P. (2010). *Artificial intelligence: a modern approach*. Pearson Education Limited, third edition.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261), 218–23.

Schnell, P. M., Tang, Q., Offen, W. W., & Carlin, B. P. (2016). A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, 72(4), 1026–1036.

Schulz, K. F. (1998). Randomized controlled trials. *Clin Obstet Gynecol*, 41(2), 245–56.

Science Daily (2013). Big data, for better or worse: 90% of world's data generated over last two years.

Scutari, M. & Ness, R. (2012). bnlearn: Bayesian network structure learning, parameter learning and inference. *R package version*, 3.

Seyednasrollah, F., Koestler, D. C., Wang, T., Piccolo, S. R., Vega, R., Greiner, R., Fuchs, C., Gofer, E., Kumar, L., Wolfinger, R. D., Kanigel Winner, K., Bare, C., Neto, E. C., Yu, T., Shen, L., Abdallah, K., Norman, T., Stolovitzky, G., Soule, H. R., Sweeney, C. J., Ryan, C. J., Scher, H. I., Sartor, O., Elo, L. L., Zhou, F. L., Guinney, J., Costello, J. C., & Prostate Cancer, D. C. C. (2017). A dream challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. *JCO Clin Cancer Inform*, 1, 1–15.

Sheng, E., Li, W., & Zhou, X. H. (2019). Estimating causal effects of treatment in rcts with provider and subject noncompliance. *Stat Med*, 38(5), 738–750.

Sun, Y. V. & Hu, Y. J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv Genet*, 93, 147–90.

Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018). Innovation at the intersection of clinical trials and real-world data science to advance patient care. *Clinical and translational science*, 11(5), 450–460.

Štajduhar, I., Dalbelo-Bašić, B., & Bogunović, N. (2009). Impact of censoring on learning bayesian networks in survival modelling. *Artificial Intelligence in Medicine*, 47(3), 199–217.

Thompson, J., Hu, J., Mudaranthakam, D. P., Streeter, D., Neums, L., Park, M., Koestler, D. C., Gajewski, B., Jensen, R., & Mayo, M. S. (2019). Relevant word order vectorization for improved natural language processing in electronic health records. *Sci Rep*, 9(1), 9253.

Thompson, J. A., Christensen, B. C., & Marsit, C. J. (2018). Methylation-to-expression feature models of breast cancer accurately predict overall survival, distant-recurrence free survival, and pathologic complete response in multiple cohorts. *Scientific reports*, 8(1), 5190–5190.

US Food and Drug Administration (2018). Framework for fda's real-world evidence program.

US Food and Drug Administration (2019). Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products guidance for industry.

Van Norman, G. A. (2016). Drugs, devices, and the fda: Part 1: An overview of approval processes for drugs. *JACC Basic Transl Sci*, 1(3), 170–179.

Vanderweele, T. J. (2011). Principal stratification–uses and limitations. *Int J Biostat*, 7(1).

VanderWeele, T. J. & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *J Causal Inference*, 1(1), 1–20.

Wang, C., Li, H., Chen, W. C., Lu, N., Tiwari, R., Xu, Y., & Yue, L. Q. (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *J Biopharm Stat*, 29(5), 731–748.

Wang, L., Tang, H., Thayanithy, V., Subramanian, S., Oberg, A. L., Cunningham, J. M., Cerhan, J. R., Steer, C. J., & Thibodeau, S. N. (2009). Gene networks and micrornas implicated in aggressive prostate cancer. *Cancer Res*, 69(24), 9490–7.

Weisberg, H. I. & Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? *Clin Trials*, 12(4), 357–64.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12(4), e0174944.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Chapter 9 - Probabilistic methods*, (pp. 335–416). Morgan Kaufmann.

Yella, J. K., Yaddanapudi, S., Wang, Y., & Jegga, A. G. (2018). Changing trends in computational drug repositioning. *Pharmaceuticals (Basel)*, 11(2).

Zarayeneh, N., Ko, E., Oh, J. H., Suh, S., Liu, C., Gao, J., Kim, D., & Kang, M. (2017). Integration of multi-omics data for integrative gene regulatory network inference. *Int J Data Min Bioinform*, 18(3), 223–239.

Zheng, C., Dai, R., Gale, R. P., & Zhang, M. J. (2020). Causal inference in randomized clinical trials. *Bone Marrow Transplant*, 55(1), 4–8.

Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M. J., Song, L., Landi, M. T., Ghosh, D., Chatterjee, N., Baladandayuthapani, V., & Zhao, H. (2017). Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci Rep*, 7(1), 16954.

Appendix A: Variable Definitions

Table A.1 shows variable definition, their data generation simulation values and Bayesian MCMC node representation.

Table A.1: Variable definition and simulation values

Variable	Description	Data generation distributions	Node (Bayesian)
i	Unit/subject index		
Z	Control (z0)	1 : 1	θ_{0S}
	Drug (z1)		θ_{1S}
X_1	Baseline HbA1c	$N(8.66, 1.09^2)$	δ_1, β_1
X_2	Baseline age	$N(58, 9^2)$	δ_2, β_2
X_3	Disease duration	$N(12, 6^2)$: Truncated (1,20)	δ_3, β_3
X_4	Baseline BMI	$N(29, 5^2)$	δ_4, β_4
X_5	Sex	$Male = Bern(0.53)$	δ_5, β_5
X_6	Concomitant drugs	$Yes = Bern(0.60)$	δ_6, β_6
D	ICE under control	Binary function	π_{01}
	ICE under drug	of X_1, X_2, X_3, X_4	π_{11}
S	Strata	Conditional categorical function	π_s
		of Z, D, X_1, X_2, X_3, X_4	
Y	Change in HbA1c	Conditional regression function	Outcome
		of Z, D, X_1	

Appendix B: Other Results Tables

Table B.1: *Scenario iii.* Heterogeneous treatment effect (b) ($\Delta_1 = -1.25,\ \Delta_2 = -0.25,\ \Delta_3 = -0.75,\ \Delta_4 = -0.05$); Weighted $\Delta = -1.012$ ($n = 1000$)

Sample Size	Estimator	Mean	SE	2.5%	50%	97.5%
50	CATE	-1.49	0.26	-2.02	-1.483	-0.992
	WATE	-1.418	0.242	-1.896	-1.416	-0.945
	ITT	-1.443	0.241	-1.914	-1.445	-0.964
	MSE_{CATE}	0.126	0.167	< 0.001	0.063	0.594
	$MSE\text{-}C_{ITT}$	0.095	0.126	< 0.001	0.049	0.444
	MSE_{WATE}	0.18	0.191	< 0.001	0.121	0.683
	$MSE\text{-}W_{ITT}$	0.197	0.198	0.001	0.142	0.713
100	CATE	-1.374	0.211	-1.787	-1.373	-0.96
	WATE	-1.424	0.196	-1.806	-1.424	-1.039
	ITT	-1.416	0.195	-1.801	-1.415	-1.031
	MSE_{CATE}	0.06	0.084	< 0.001	0.028	0.293
	$MSE\text{-}C_{ITT}$	0.066	0.085	< 0.001	0.034	0.304
	MSE_{WATE}	0.231	0.181	0.005	0.193	0.674
	$MSE\text{-}W_{ITT}$	0.224	0.177	0.004	0.185	0.666
200	CATE	-1.263	0.176	-1.596	-1.268	-0.902
	WATE	-1.131	0.126	-1.378	-1.131	-0.882
	ITT	-1.135	0.131	-1.394	-1.135	-0.878
	MSE_{CATE}	0.031	0.046	< 0.001	0.014	0.16
	$MSE\text{-}C_{ITT}$	0.03	0.039	< 0.001	0.016	0.139
	MSE_{WATE}	0.04	0.045	< 0.001	0.025	0.161
	$MSE\text{-}W_{ITT}$	0.042	0.049	< 0.001	0.026	0.175
400	CATE	-1.28	0.124	-1.521	-1.28	-1.038
	WATE	-1.153	0.085	-1.318	-1.152	-0.987
	ITT	-1.152	0.088	-1.324	-1.151	-0.98
	MSE_{CATE}	0.016	0.023	< 0.001	0.007	0.081
	$MSE\text{-}C_{ITT}$	0.017	0.021	< 0.001	0.01	0.073
	MSE_{WATE}	0.03	0.028	< 0.001	0.023	0.101
	$MSE\text{-}W_{ITT}$	0.03	0.029	< 0.001	0.023	0.104
1000	CATE	-1.208	0.079	-1.36	-1.21	-1.051
	WATE	-1.182	0.055	-1.29	-1.182	-1.073
	ITT	-1.181	0.057	-1.292	-1.181	-1.07
	MSE_{CATE}	0.008	0.011	< 0.001	0.004	0.04
	$MSE\text{-}C_{ITT}$	0.008	0.009	< 0.001	0.005	0.032
	MSE_{WATE}	0.032	0.019	0.004	0.029	0.077
	$MSE\text{-}W_{ITT}$	0.032	0.02	0.003	0.029	0.079

Table B.2: *Scenario iv.* Reversed treatment effect ($\Delta_1 = -1.25,\ \Delta_2 = -2.50,\ \Delta_3 = 3.00,\ \Delta_4 = 0.00$); Weighted $\Delta = -1.172 (n = 1000)$

Sample Size	Estimator	Mean	SE	2.5%	50%	97.5%
50	CATE	-0.946	0.327	-1.608	-0.942	-0.315
	WATE	-0.897	0.313	-1.516	-0.896	-0.279
	ITT	-0.909	0.311	-1.518	-0.911	-0.288
	MSE_{CATE}	0.199	0.248	< 0.001	0.11	0.877
	$MSE\text{-}C_{ITT}$	0.214	0.26	< 0.001	0.123	0.925
	MSE_{WATE}	0.098	0.144	< 0.001	0.043	0.508
	$MSE\text{-}W_{ITT}$	0.097	0.145	< 0.001	0.043	0.502
100	CATE	-1.576	0.235	-2.031	-1.578	-1.105
	WATE	-1.618	0.215	-2.041	-1.618	-1.195
	ITT	-1.589	0.247	-2.077	-1.589	-1.102
	MSE_{CATE}	0.162	0.17	< 0.001	0.11	0.61
	$MSE\text{-}C_{ITT}$	0.176	0.189	< 0.001	0.116	0.683
	MSE_{WATE}	0.149	0.153	< 0.001	0.104	0.553
	$MSE\text{-}W_{ITT}$	0.146	0.169	< 0.001	0.088	0.607
200	CATE	-1.199	0.202	-1.601	-1.194	-0.816
	WATE	-0.98	0.141	-1.256	-0.98	-0.701
	ITT	-0.961	0.173	-1.302	-0.961	-0.623
	MSE_{CATE}	0.043	0.058	< 0.001	0.021	0.209
	$MSE\text{-}C_{ITT}$	0.113	0.109	0.001	0.084	0.394
	MSE_{WATE}	0.02	0.029	< 0.001	0.009	0.103
	$MSE\text{-}W_{ITT}$	0.031	0.045	< 0.001	0.014	0.157
400	CATE	-1.275	0.11	-1.497	-1.273	-1.067
	WATE	-1.322	0.088	-1.495	-1.322	-1.149
	ITT	-1.294	0.108	-1.505	-1.293	-1.082
	MSE_{CATE}	0.013	0.02	< 0.001	0.006	0.066
	$MSE\text{-}C_{ITT}$	0.014	0.019	< 0.001	0.006	0.068
	MSE_{WATE}	0.045	0.036	0.001	0.037	0.134
	$MSE\text{-}W_{ITT}$	0.039	0.039	< 0.001	0.027	0.142
1000	CATE	-1.332	0.069	-1.467	-1.331	-1.2
	WATE	-1.275	0.056	-1.386	-1.275	-1.164
	ITT	-1.278	0.072	-1.419	-1.278	-1.138
	MSE_{CATE}	0.011	0.013	< 0.001	0.007	0.047
	$MSE\text{-}C_{ITT}$	0.006	0.008	< 0.001	0.003	0.03
	MSE_{WATE}	0.014	0.012	< 0.001	0.011	0.046
	$MSE\text{-}W_{ITT}$	0.016	0.017	< 0.001	0.011	0.061

Table B.3: *Scenario v.* No treatment effect ($\Delta_s = 0.00$); Weighted $\Delta = 0.00 (n = 1000)$

Sample Size	Estimator	Mean	SE	2.5%	50%	97.5%
50	CATE	-0.365	0.248	-0.859	-0.363	0.117
	WATE	-0.34	0.238	-0.807	-0.339	0.127
	ITT	-0.347	0.238	-0.812	-0.349	0.127
	MSE_{CATE}	0.195	0.206	0.001	0.133	0.737
	$MSE\text{-}C_{ITT}$	0.177	0.184	< 0.001	0.123	0.66
	MSE_{WATE}	0.172	0.182	< 0.001	0.116	0.651
	$MSE\text{-}W_{ITT}$	0.177	0.184	< 0.001	0.123	0.66
100	CATE	-0.17	0.207	-0.577	-0.17	0.237
	WATE	-0.246	0.192	-0.623	-0.246	0.131
	ITT	-0.231	0.192	-0.61	-0.23	0.148
	MSE_{CATE}	0.071	0.093	< 0.001	0.036	0.334
	$MSE\text{-}C_{ITT}$	0.09	0.103	< 0.001	0.055	0.372
	MSE_{WATE}	0.097	0.108	< 0.001	0.062	0.388
	$MSE\text{-}W_{ITT}$	0.09	0.103	< 0.001	0.055	0.372
200	CATE	-0.066	0.179	-0.401	-0.072	0.306
	WATE	-0.002	0.129	-0.253	-0.002	0.251
	ITT	-0.005	0.131	-0.263	-0.005	0.252
	MSE_{CATE}	0.037	0.05	< 0.001	0.017	0.178
	$MSE\text{-}C_{ITT}$	0.017	0.025	< 0.001	0.008	0.087
	MSE_{WATE}	0.017	0.024	< 0.001	0.007	0.083
	$MSE\text{-}W_{ITT}$	0.017	0.025	< 0.001	0.008	0.087
400	CATE	-0.057	0.126	-0.298	-0.058	0.192
	WATE	0.025	0.085	-0.142	0.024	0.191
	ITT	0.026	0.088	-0.145	0.027	0.197
	MSE_{CATE}	0.019	0.026	< 0.001	0.009	0.092
	$MSE\text{-}C_{ITT}$	0.008	0.012	< 0.001	0.004	0.042
	MSE_{WATE}	0.008	0.011	< 0.001	0.004	0.039
	$MSE\text{-}W_{ITT}$	0.008	0.012	< 0.001	0.004	0.042
1000	CATE	0.005	0.08	-0.147	0.003	0.166
	WATE	-0.023	0.055	-0.129	-0.022	0.084
	ITT	-0.021	0.056	-0.131	-0.021	0.087
	MSE_{CATE}	0.006	0.009	< 0.001	0.003	0.032
	$MSE\text{-}C_{ITT}$	0.004	0.005	< 0.001	0.002	0.018
	MSE_{WATE}	0.003	0.005	< 0.001	0.002	0.017
	$MSE\text{-}W_{ITT}$	0.004	0.005	< 0.001	0.002	0.018

Appendix C: JAGs Code

The JAGs code for the principal stratification with strata predictive covariates is provided.

```
#=================================================================
# Model Section
#=================================================================

model {
for (i in 1:N)
{
#=================================================================
# RANDOM/COVARIATE PREDICTION OF LATENT STRATA
#=================================================================
# P01 is probability of event in control group
# P01 (in Rx) are counterfactual predicted probabilities of event under treatment
# P11 is probability of event in treatment group
# P11 (in Rx) are counterfactual predicted probabilities of event under control

Dz0[i] ~ dbern(p01[i])
logit(p01[i]) <- delta0[1]*step(z0[i]-1) + delta1[1]*X1[i]*step(z0[i]-1) + delta2[1]*X2[i
    ]*step(z0[i]-1) + delta3[1]*X3[i]*step(z0[i]-1) + delta5[1]*X5[i]*step(z0[i]-1)

Dz1[i] ~ dbern(p11[i])
logit(p11[i]) <- delta0[2]*step(z1[i]-1) + delta1[2]*X1[i]*step(z1[i]-1) + delta2[2]*X2[i
    ]*step(z1[i]-1) + delta3[2]*X3[i]*step(z1[i]-1) + delta5[2]*X5[i]*step(z1[i]-1)

# Counterfactual posterior predictive values on the control for those on the treatment
p1z0.ppred[i] <- 1/(1+exp(-(delta0[1] + delta1[1]*X1[i] + delta2[1]*X2[i] + delta3[1]*X3[
    i] + delta5[1]*X5[i])))
```

```
# Counterfactual posterior predictive values on the treatment for those on the control
p1z1.ppred[i] <- 1/(1+exp(-(delta0[2] + delta1[2]*X1[i] + delta2[2]*X2[i] + delta3[2]*X3[
    i] + delta5[2]*X5[i])))

p[i,1] <- (1-p1z1.ppred[i])*(1-D[i])*(1-Z[i]) + (1-p1z0.ppred[i])*(1-D[i])*(Z[i])
p[i,2] <- (p1z1.ppred[i])*(1-D[i])*(1-Z[i]) + (1-p1z0.ppred[i])*(D[i])*(Z[i])
p[i,3] <- (1-p1z1.ppred[i])*(D[i])*(1-Z[i]) + (p1z0.ppred[i])*(1-D[i])*(Z[i])
p[i,4] <- (p1z1.ppred[i])*(D[i])*(1-Z[i]) + (p1z0.ppred[i])*(D[i])*(Z[i])

S[i] ~ dcat(p[i,1:4])

#=================================================================
# MODELING OUTCOME IN EACH STRATUM
#=================================================================
# Assumes same variance across all strata
Y[i] ~ dnorm(mu[i], prec)
# Assumes same effects of covariate on outcomes across the four strata
mu[i] <- theta0[S[i]]*z0[i]+theta1[S[i]]*z1[i] + beta1*X1[i] + beta2*X2[i]+ beta3*X3[i]+
    beta4*X4[i] +beta5*X5[i]
}
#=================================================================
# PRIORS
#=================================================================
#####################################################################
# Define prior for within strata regression parameters
#####################################################################
# Treatment effect priors
#-------------------------------------------------
for (r in 1:4) {
       theta0[r] ~ dnorm(0, 1.0e-4)
       theta1[r] ~ dnorm(0, 1.0e-4)
}
```

```
# Non-informative prior for precision
prec ~ dgamma(0.001, 0.001)
sigma <- 1/sqrt(prec)

# Latent strata predictor parameter priors
#----------------------------------------

for (d in 1:2) {
       delta0[d] ~ dnorm(0, 1.0e-4)
       delta1[d] ~ dnorm(0, 1.0e-4)
       delta2[d] ~ dnorm(0, 1.0e-4)
       delta3[d] ~ dnorm(0, 1.0e-4)
       delta5[d] ~ dnorm(0, 1.0e-4)
}

# Priors for regression parameters
#----------------------------------------
beta1 ~ dnorm(0, 1.0e-4)
beta2 ~ dnorm(0, 1.0e-4)
beta3 ~ dnorm(0, 1.0e-4)
beta4 ~ dnorm(0, 1.0e-4)
beta5 ~ dnorm(0, 1.0e-4)

#================================================================
# CALCULATED QUANTITIES
#================================================================

# Estimate of treatment effect in each strata as a difference between drug and placebo
for(ITTS in 1:4){
       trt_diff[ITTS] <- theta1[ITTS] - theta0[ITTS]
       }
```

```
# Estimated strata proportions based on latent membership assignment.
for(pr in 1:4){
       prp.strata[pr] <- mean(equals(S, pr))
       }

# Estimated proportions by treatment group and strata

for(tr in 1:2){
       for (st in 1:4){
              prop.ZS[tr, st] <- mean(equals(Z,(tr-1)) && equals(S,st))
              }
       }

#================================================================
# CALCULATED QUANTITIES
# Both CATE and WATE are known quantities from simulation
#================================================================
# PERFOMANCE ASSESSMENT METRICS
# Causal Average Treatment Effect (Stratum 1)

# Mean Square Error (MSE)
MSEc <- pow((trt_diff[1] - CATE), 2)

# Weighted Causal Average Treatment Effect (Stratum 1, 2, 3) : This is due to the imposed
     missingness after ICE

W.CATE <- ((((theta1[1]*prop.ZS[2,1])/(prop.ZS[2,1] + prop.ZS[2,3]))+ ((theta1[3]*prop.ZS
    [2,3])/(prop.ZS[2,1] + prop.ZS[2,3]))) -
(((theta0[1]*prop.ZS[1,1])/(prop.ZS[1,1] + prop.ZS[1,2]))+ ((theta0[2]*prop.ZS[1,2])/(
    prop.ZS[1,1] + prop.ZS[1,2])))))

MSEw <- pow((W.CATE - WATE), 2)
}
```

```
#============================================================
# OTHER SUPPLEMENTARY ELEMENTS OF THE CODE
# Assuming real or simulated data with the following variables
#============================================================

# Y  Continuous outcomes: Simulated change in HbA1C from baseline
# Z  Randomization
# D  Intercurrent event status
# X1 Baseline HbA1C
# X2 Baseline Age
# X3  Baseline Disease Duration
# X4  Sex
# X5  Race
# N  Sample Size

# Other generated variables
z0 <- ifelse(Z==0, 1, 0)  # Indicator variable of control assignment =1 if control, 0
    otherwise
z1 <- ifelse(Z==1, 1, 0)  # Indicator variable of treatment assignment =1 if treatment, 0
     otherwise
Dz0 <- ifelse(Z==0, D, NA)  # Indicator variable of ICE under control; =1 if control, 0
    otherwise
Dz1 <- ifelse(Z==1, D, NA)  # Indicator variable of ICE under treatment; =1 if control, 0
     otherwise
S <- c(rep(NA,N))    # Assign missing strata for all subjects

#============================================================
# TRUE Values of treatment effects
#============================================================
# Empirical==TRUE implies the true effects estimated from the data
# betaxmat: Matrix with simulation values; treatment and covariate effects
# mn_sk_zt: Calculated empirical mean of outcomes stratum k, treatment t
```

```
# Observed Weighted Average Treatment Effect
CATE <- ifelse(empirical==TRUE, mn_s1_z1 - mn_s1_z0, betaxmat[1,2] - betaxmat[1,1])

WATE <- ifelse(empirical==TRUE,
       sum(c(p1*(mn_s1_z1 - mn_s1_z0), p2*(mn_s2_z1 - mn_s2_z0),
       p3*(mn_s3_z1 - mn_s3_z0), p4*(mn_s4_z1 - mn_s4_z0)), na.rm = TRUE),
       sum(c(p1*(betaxmat[1,2] - betaxmat[1,1]), p2*(betaxmat[2,2] - betaxmat[2,1]),p3*(
           betaxmat[3,2] - betaxmat[3,1]), p4*(betaxmat[4,2] - betaxmat[4,1])), na.rm =
           TRUE))

#=======================================================

# Define list in standard mode used in the model
model.data <- list("Y", "Z", "z0", "z1", "D", "Dz0", "Dz1",
              "X1", "X2", "X3", "X4", "X5", "N", "S", "CATE", "WATE")

# INITIAL VALUES########################################
# Initializing parameter values
init.val <- function()
{
list( theta0=rnorm(4, 0, 2),
       theta1=rnorm(4, 0, 2),
       beta1=rnorm(1, 0, 2),
       beta2=rnorm(1, 0, 2),
       beta3=rnorm(1, 0, 2),
       beta4=rnorm(1, 0, 2),
       beta5=rnorm(1, 0, 2),
       prec = 1)
}
```

```
# Compile and output results based on predefined functions and values
output <- jags(data = model.data,
               inits = init.val,
               parameters.to.save = parameters,
               model.file = modfile,
               n.chains = n.chains,
               n.iter = n.iter,
               n.burnin = n.burn.in,
               jags.seed = jagseed,
               n.thin = nthin )
```