New Statistical Insights to Precision Medicine, from Targeted Treatment Development to Individualized Tailoring Recommendation

by

Yue Wei

BS, Peking University, 2012

MS, University of Illinois, 2016

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yue Wei

It was defended on

July 26th 2021

and approved by

Ying Ding, PhD, Associate Professor of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

Chaeryon Kang, PhD, Assistant Professor of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

Jong H. Jeong, PhD, Professor and Interim Chair of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

Chung-Chou H. Chang, PhD, Professor of Medicine, Biostatistics, and Clinical and Translational Science

School of Medicine, Graduate School of Public Health, University of Pittsburgh

Yu Cheng, PhD, Professor of Statistics,

School of Arts and Sciences, University of Pittsburgh

New Statistical Insights to Precision Medicine, from Targeted Treatment Development to Individualized Tailoring Recommendation

Yue Wei, PhD

University of Pittsburgh, 2021

Abstract

There has been increasing interest in discovering precision medicine in current drug development. One aspect of precision medicine is to develop new therapies that target a subgroup of patients with enhanced treatment efficacy through clinical trials. Another aspect is to tailor existing therapies to each patient so that everyone can get the most "suitable" treatment. Motivated by analyzing the Age-Related Eye Disease Study (AREDS) data, this dissertation proposes new statistical methods to address issues in both aspects.

In the first part, I propose a novel multiple-testing-based approach to simultaneously identify and infer subgroups with enhanced treatment efficacy. Specifically, I formulate the null hypotheses through contrasts and construct their simultaneous confidence intervals, which control both within- and across-marker multiplicity. Two types of outcomes are considered: survival and binary endpoints. Extensive simulations are conducted to evaluate the method performance and provide practical guidance. The method is then applied to AREDS data to assess the efficacy of antioxidants and zinc in delaying AMD progression. I further validate the findings in AREDS2, by discovering consistent differential treatment responses in subgroups identified from AREDS.

In the second part, I develop machine-learning-based approaches to estimate individual treatment effects (ITE) so that individualized tailoring recommendation can be provided. Specifically, I implement random survival forest, Bayesian accelerated failure time model, and Cox-based deep neural network survival model under the framework of meta-algorithms: T-learner and X-learner, to accurately estimate ITEs with survival outcomes. Treatment recommendation rule is provided based on patient's ITE estimate and then evaluated by various performance metrics. I investigate the merits of the proposed methods with comprehensive

simulation studies and apply them on AREDS data. Finally, the Boruta algorithm is applied to identify top variables that contribute to the treatment recommendation rule.

Public health significance: This dissertation addresses two precision medicine research questions: (1) targeted treatment development, i.e., whether there exists subgroup of patients with beneficial treatment efficacy; (2) tailoring existing therapies through ITE estimation. It has the potential to significantly improve the current practice in analyzing treatment effects, and thus to increase the success of modern drug development and precision medicine research.

Table of Contents

List of Tables

List of Figures

Preface

Firstly, I would like to thank my advisor, Dr. Ying Ding, who is always supportive and helps me make the right decisions. It's my great honor and pleasure to have her guide me during my PhD life at Pitt. She provides many opportunities with me to build my profile and explore all kinds of career possibilities. In addition, she always patiently helped me revise the manuscripts, gave advice on how to efficiently communicate with collaborators, and guided me to improve my presentation skills. These are all essential elements to achieve success in both academic and industrial careers. Words are powerless to express my gratitude to her. She is the best! Secondly, I would like to thank my co-advisor, Dr. Chaeryon Kang, who provided insightful suggestions to my dissertation, especially the second part about CATE estimation and choosing the best treatment for each individual (Chapter 3). She is always patient to answer my questions and willing to share her experience with me. I would also like to thank Dr. Jong Jeong, Dr. Chung-Chou H. Chang, and Dr. Yu Cheng, to serve as my committee members. They all brought up great points regarding my dissertation and shed light on the directions for my future research topics.

My special thanks go to Dr. Wei Chen, who kindly offer me the opportunity to join his group meeting where I gained genomic and genetic knowledge, together with the cutting-edge statistical methods to analyze single-cell experiments. I also want to thank Dr. Robert A. Sweet, Dr. Julia Kofler, Dr. Mattew L. McDonald, Dr. Brandon Mckinney, and Dr. Sohail Husain. I did my GSR work with them under the guidance of Dr. Ding. They are all outstanding scientists who shared their expertise in clinical and biomedical areas. Working with them helps me understand how statistics can assist solving the real world problem. Additional thanks go to the former and current group members, including but not limited to Dr. Tao Sun, Xinjun Wang, Na Bo and Lang Zeng, who have helped me in my research projects.

Finally, and most importantly, huge thank you to my families, for their deepest love and encouragement. I would like to thank my husband, Yuping Wang, for being patient and listening, and always supportive. A special thank you goes to my mom, Xianbo Cui,

for coming to the US from the other side of the planet, overcoming all difficulties including quarantining herself in the third country for 14 days, just to support me by taking care of the little baby so that I can focus on my research. Without her, this would have never been accomplished. Thank you and I love you, Mom.

1.0 Introduction

1.1 Overview

Traditional medical treatments are often designed for the "average patient" as a "one-fits-all" approach, which may benefit some patients but not everyone. Precision medicine, as an innovative approach for disease treatment and prevention, takes into account individual variability in genes, environments and lifestyles. Motivated by analyzing the Age-Related Eye Disease Study (AREDS) data, a large randomized clinical trial (RCT) to study the efficacy of nutritional supplements in delaying the progression of age-related macular degeneration (AMD), this dissertation proposes new statistical insights to precision medicine, from targeted treatment development based on genetic factors of a patient, to individualized tailoring recommendation based on the heterogeneous treatment effects estimation.

In the rest of this Chapter, I will start by introducing the basic concepts, such as time-to-event data and binary data, in Sections 1.2 and 1.3. In the following, the concept of "logic-respecting" treatment efficacy measurements and the commonly used ones will be introduced for both types of outcomes in Section 1.4. Lastly several meta-algorithms to estimate the conditional average treatment effect (CATE) will be discussed, followed by three machine learning models for survival outcomes in Section 1.5.

1.2 Time-to-event data

Time-to-event data is a special type of data that describes time to a well defined endpoint of interest (e.g., death, heart attack, onset of a pandemic, and remission of cancer). It is also known as "survival data", even though the outcome is not always death. There are unique features of time-to-event variables. First of all, time-to-event variables are always positive and the distribution can be skewed. Secondly, complete data is not always available. For example, death could certainly happen after the study ends and the observed time when

patients exit the study is not the actual time to death. Instead, it is called censored time.

There are three types of censoring. The most common one is the right censoring and occurs when a participant does not have the event of interest during the study and thus their last observed time is less than their actual time to event. This can occur when a participant drops out before the study ends or when a participant is event free by the end of the study. Another type of censoring is the left censoring. It happens when a recruited participant already has the event of interest prior entry of the study but the time of developing the event is unknown. The last type is interval censoring when the time to event is known only to lie within an interval instead of being observed exactly. Truncation often adds complexity to the data analysis and it is different from the censoring. Truncation is due to sampling bias that only individuals satisfying specific conditions could be recruited to the study.

Because of the unique features of time-to-event data, the analysis of such data, or survival analysis, requires different statistical techniques.

1.2.1 Models for time-to-event data

The survival function of event time (T) is defined as the probability that T is greater than a given time t:

$$S(t) = P(T > t),\ 0 < t < \infty.$$

When T is absolutely continuous, we have a one-to-one relationship between the survival function $S(t)$, density function $f(t)$, hazard function $\lambda(t)$, and cumulative hazard function $\Lambda(t)$, expressed as:

$$S(t) = e^{-\Lambda(t)},$$

where

$$\Lambda(t) = \int_0^t \lambda(s)ds,\ \lambda(t) = f(t)/S(t).$$

Survival functions and (cumulative) hazard functions are more commonly used for modeling the survival time T than the density function. Like analysis of other types of outcomes, a key aspect of survival analysis is to understand the relationship between covariates and the survival function through regression models.

1.2.1.1 Cox proportional hazards model

The most popular regression model of survival analysis is the Cox proportional hazard (CoxPH) model (Cox, 1972a). The hazard function can be expressed as:

$$\lambda(t; X) = \lambda_0(t) \exp(\beta X), \tag{1.2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, X is a vector of covariates, and β is a vector of covariate coefficient parameters. The method does not assume the baseline hazard function, but it has a key assumption that the effects of the predictor variables upon survival are constant over time and are additive in one scale. The model is interpreted as the ratio of hazard functions is a constant (independent of time t) between two subjects with different X. The coefficients in a Cox regression relate to hazard; a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated.

Integrate both sides of equation 1.2.1 from 0 to t to obtain the cumulative hazards:

$$\Lambda(t; X) = \Lambda_0(t) \exp(\beta X),$$

which are also proportional. Further, we can derive the survivor function as:

$$S(t; X) = e^{-\Lambda_0(t) \exp(\beta X)} = \{S_0(t)\}^{\exp(\beta X)},$$

where $S_0(t) = \exp(-\Lambda_0(t))$ is a baseline survival function. Thus, the effect of the covariate values X on the survivor function is to raise it to a power given by the relative risk $\exp(\beta X)$.

The PH model is the most popular regression model in survival analysis due to the partial likelihood approach for right-censored failure time data (Cox, 1972a). The approach is simple and efficient because the partial likelihood only involves the finite-dimensional β parameter without the nuisance infinite-dimensional $\lambda_0(t)$. The resulting β estimate is asymptotically equivalent to that obtained from the full likelihood.

1.2.1.2 Accelerated failure time model

Another commonly used regression model for survival analysis is the accelerated failure time (AFT) model (Wei, 1992), which proposes the following relationship between covariates and $\log T$:

$$\log T = \mu + \alpha X + \sigma W,$$

where β is a vector of coefficients, and W follows an unspecific distribution. The above framework describes a general class of models. Depending on the distribution for W, we will obtain different models, but all will have the same general structure. Common distributions used in AFT models includes normal distribution (also known as log-normal model), standard logistic distribution (log-logistic model), and 2-parameter extreme value distribution (Weibull model). Notice that Weibull model is the only model that satisfies both AFT and PH. Given transformations

$$\begin{aligned} \gamma &= \frac{1}{\sigma}, \\ \lambda &= \exp(-\frac{\mu}{\sigma}), \\ \beta &= -\frac{\alpha}{\sigma}, \end{aligned}$$

we have a Weibull model with baseline hazard of

$$h(t|x) = (\gamma\lambda t^{\gamma-1}) \exp(\beta x).$$

Note that different approaches result in different coefficient parameters in a Weibull regression model. The signs of α and β are opposite. Special caution is needed for interpretation of the coefficient parameters.

1.3 Binary data

Binary data refer to those that can take only one of the two values, such as true or false questions, mortality (dead or alive), and flipping a coin (head or tail). Binary variables have a variety of applications including but not limited to medical diagnoses, facial recognition, and decision trees. One of the popular classification example using neural networks is to determine whether an image is a cat or a dog. For biomedical research such as clinical trials, binary outcomes serve as a key measure to compare the treatment effects on disease status. As for all types of outcomes, the analysis of binary outcomes provides unique aspect of scientific practice, and different statistical considerations are needed as compared to other types of outcomes such as continuous and time-to-event variables.

1.3.1 Models for binary data

Assume Y follows Bernoulli distribution with $P(Y = 1) = \pi$, we will consider models for π, which can depend on explanatory variables (i.e., $x_1, x_2, \ldots, x_p$).

$$\pi(x) = P(Y = 1|x_1, x_2, \ldots, x_p).$$

Using the generalized linear model,

$$g(\pi(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Different link function g can result in different models.

1.3.1.1 Logistic model

Logistic regression uses the logit link to explain the relationship between π and explanatory variables.

$$g(\pi(x)) = \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Whereas $0 \leq \pi \leq 1$, the range for $\text{logit}(\pi)$ is all real numbers. However, with the transformation using logit link, the interpretation of the coefficients estimates is on the scale of odds ratio (OR).

1.3.1.2 Log-binomial model

Similar to the logistic model, the error term of Y is assumed to follow a Binomial distribution. The different assumption is that the link function is now a log function.

$$g(\pi(x)) = \log(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Since the interpretation of odds ratio is not straightforward and sometimes hard to understand, some researchers prefer using relative risk (RR) instead. Although for rare events, OR may serve as a good approximation for RR, when events are common, OR tends to overestimate the risk ratio. The log-binomial model provides a natural interpretation on the RR scale and thus is a better alternative for the analysis of cross-sectional studies with binary outcomes (Barros and Hirakata, 2003). When the link function is misspecified or when the probability distribution of the response variable was truncated, the log-binomial model tends to provide biased point estimates. A modified Poisson model introduced by Zou (2003) is generally preferable (Chen et al., 2018).

1.4 Logic-respecting treatment efficacy measures

Typically the goal of a RCT is to compare the new treatment (denoted by Rx) with a control such as a placebo or a standard of care (denoted by C). The "relative" treatment effect between Rx and C describes the treatment efficacy. When heterogeneity exists in the population, the measurement of treatment efficacy in subgroups and in combination of subgroups is required. Assume a marker separates the population into "marker positive" group (g^+) and "marker negative" (g^-) group. In targetd treatment development process, researchers care about not only the treatment efficacy in g^+ or g^-, but also that in the combined group $\{g^+, g^-\}$. Formally described in Lin et al. (2019), a "logic-respecting" treatment efficacy measure requires that the treatment efficacy in the combined group should be within the range defined by the treatment efficacy for the two individual groups. If we use μ to denote treatment efficacy, then a logic-respecting treatment efficacy measure should

satisfy $\mu_{\{g^+,g^-\}} \in [\mu_{g^-}, \mu_{g^+}]$, assuming $\mu_{g^-} < \mu_{g^+}$. Although it seems trivial, it has not been fully recognized and some commonly used efficacy measures are not logic-respecting for mixture populations. We will demonstrate this issue with examples in time-to-event and binary outcomes.

1.4.1 Efficacy measures for time-to-event outcomes

Time-to-event outcomes, also known as survival outcomes, are commonly used in clinical trials, especially in oncology studies. These outcomes take account of both whether the event occurs and the timing of the event. The most widely used model to analyze time-to-event data is the CoxPH regression, where the hazard ratio (between Rx and C) is obtained from the coefficient estimates, and has been a commonly used treatment efficacy measure. However, as shown in Ding et al. (2016), the HR is not a proper efficacy measure to use when the population is a mixture of subgroups. This is because the overall population typically does not have a constant HR. In fact, the HR of the mixture population is usually a complex function of time, with values at some time points outside the range of [HR_{g^-},HR_{g^+}]. This is because $\text{HR}_{\{g^+,g^-\}}$ can not be expressed as a weighted combination of HR_{g^-} and HR_{g^+}. The combination can only be made on density or cumulative density functions, not on the hazard ratio scale. For example, we can generate data from a Weibull distribution where $\text{HR}_{g^-} = \exp(0.3) = 1.35$, and $\text{HR}_{+} = \exp(-0.4) = 0.67$ with an equal prevalence of the two subgroups. However, the true $\text{HR}_{\{g^+,g^-\}}$ is a smooth function of time and goes below 0.67 when t is large (Figure 1.4.1). Thus using HR as the efficacy measure can lead to paradoxical findings in patient targeting.

Other commonly accepted treatment efficacy measures under consideration include the ratio or difference of: (1) survival probability at a specific timepoint, (2) mean (restricted) survival time, and (3) quantile survival time. Ding et al. (2016) demonstrated that the ratio or difference of mean or median survival times is logic-respecting. In addition, they have more direct clinical interpretations compared to HR. In this dissertation, we consider using the ratio of quantile survival times as treatment efficacy measure in the first part of Chapter 2 and the difference of survival probability as treatment efficacy measures in Chapter 3.

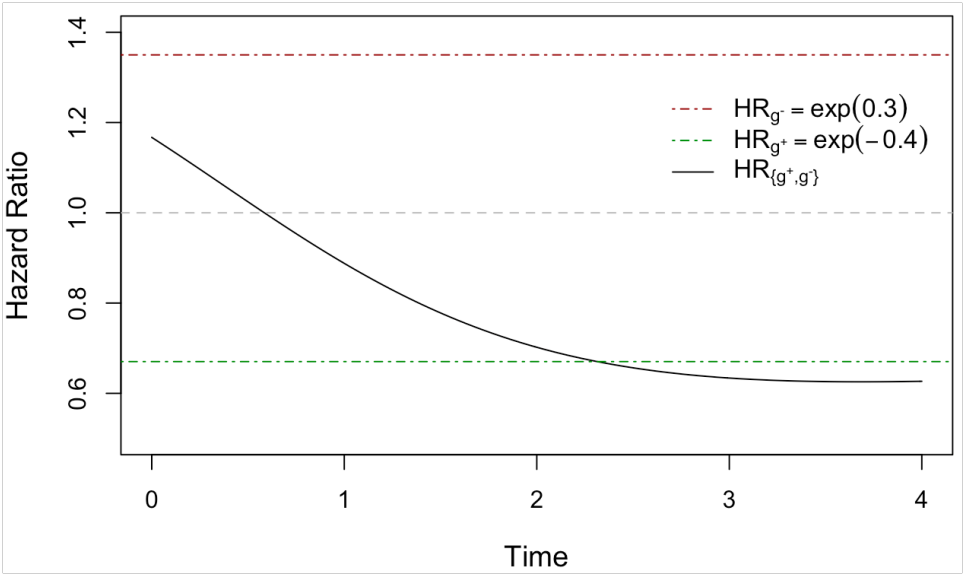

Figure 1.4.1: The plot of HR for g^-, g^+, and $\{g^+, g^-\}$.

1.4.2 Efficacy measures for binary outcomes

As described in Section 1.3, binary outcomes are often modeled using logistic regression or log-binomial models, where the OR or RR (between Rx and C) are natural choices for the efficacy measures for binary case. However, as demonstrated in Lin et al. (2019), OR is not logic-respecting, as we show below.

Table 1.4.1 gives a hypothetical example for responding or non-responding probabilities in each g^+ and g^- subgroup and the overall $\{g^+, g^-\}$. We can calculate the OR for each group and the all-comers as follows,

$$OR_{g^+} = \frac{750 \times 500}{500 \times 250} = 3,\ OR_{g^-} = \frac{250 \times 900}{100 \times 750} = 3,\ OR_{\{g^+,g^-\}} = \frac{1,000 \times 1,400}{600 \times 1,000} = \frac{7}{3}.$$

Therefore, OR is not logic-respecting since it leads to paradoxical conclusions. For example, assume OR > 2.5 indicates the new treatment is more efficacious. Then in the g^+ and g^- group individually, the Rx is clinically more efficacious while in the combined $\{g^+, g^-\}$ group it is not.

Table 1.4.1: An example of responding (R) and non-responding (NR) probabilities given Rx and C, in g^+ and g^- subgroups and the all-comers $\{g^+, g^-\}$ population

	g^+ subpopulation				g^- subpopulation				population		
	R	NR			R	NR			R	NR	
Rx	750	250		+	250	750		=	1,000	1,000	
C	500	500		+	100	900		=	600	1,400	
			1/2				1/2				1

The other commonly used efficacy measure RR, has been shown to be logic-respecting in Lin et al. (2019). Back to the previous example in Table 1.4.1,

$$RR_{g^+} = \frac{\frac{750}{750+250}}{\frac{500}{500+500}} = \frac{3}{2},\ RR_{g^-} = \frac{\frac{250}{250+750}}{\frac{100}{100+900}} = \frac{5}{2},\ RR_{\{g^+,g^-\}} = \frac{\frac{1{,}000}{1{,}000+1{,}000}}{\frac{600}{600+1{,}400}} = \frac{5}{3}.$$

Now the RR in the combined group is within the range determined by RR_{g^+} and RR_{g^-}. Rigorous proof of the property can be found in Lin et al. (2019). In the second part of Chapter 2 where our interest is on the 10-year progression status of late-AMD, the RR is used as the efficacy measure.

1.5 Meta-algorithms and machine learning models for survival data

1.5.1 Framework and definitions

Let $(Y_i(0), Y_i(1), \boldsymbol{X_i}, Z_i)$ denote the dataset of patient i under Neyman-Rubin potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990), where $\boldsymbol{X_i}$ is a $p-$dimensional covariate matrix, $Z_i \in \{0, 1\}$ is the treatment indicator, and $Y_i(0), Y_i(1)$ are the potential outcomes when i is assigned to the control group and treatment group. The causal effect of the treatment on a new patient i with the feature vector $\boldsymbol{X_i}$ can be estimated by the

individual treatment effect (ITE), which is defined as $Y_i(1) - Y_i(0)$. However, the counterfactual outcomes of the same individual cannot be obtained simultaneously. The conditional average treatment effect (CATE) can then be used to estimate the causal effect which is defined as follows (Rubin, 2005):

$$\tau(X) = E[Y_i(1) - Y_i(0)|\boldsymbol{X_i} = \boldsymbol{X}]. \tag{1.5.1}$$

Kunzel et al. (2018) has shown that the best estimate for the CATE is also the best estimate for the ITE.

To aid the estimation of CATE, the following three assumptions are needed:
textbfAssumption 1 (Consistency)

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

Assumption 2 (Unconfoundedness)

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|\boldsymbol{X_i}$$

Assumption 3 (Population Overlap)

$$P(Z_i = 1|\boldsymbol{X_i} = \boldsymbol{x_i}) \in (0, 1)$$

The consistency assumption implies that the actual observed outcome for an individual is the outcome under his or her observed exposure history. The unconfoundedness assumption requires the treatment assignment to be independent of the potential outcomes given covariates sets, which rules out the existence of unobserved factors that affect treatment choice and are also correlated with the outcomes. The population overlap assumption describes that for each value of covariate set, there is a positive probability of being assigned to both treatment and control arms, or equivalently, there is sufficient overlap in the characteristics of treated and untreated patients for adequate matches.

1.5.2 Meta-algorithms

Kunzel et al. (2018) formally defined a series of meta-algorithms (or meta-learners) to estimate CATE which take advantage of machine learning or regression estimates in a specific manner where the base learners can be any form. It adds flexibility to leverage different prior information and can be easily adapted to various types of data.

1.5.2.1 S-learner

The S-learner takes a single prediction model where the treatment indicator is included as a feature similar to all of the other covariates. In that case, the training set is defined as $\{(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)\}$. The estimated response function is then

$$\mu(x, z) = E[Y^{obs}|X = x, Z = z],$$

using any base learner on the entire dataset. Denote the estimate of the response as $\hat{\mu}$. The CATE estimate is given by

$$\hat{\theta}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0). \tag{1.5.2}$$

1.5.2.2 T-learner

The T-learner consists two steps to estimate the response function for patients assigned to treatment and control groups. First, the treatment response function,

$$\mu_1(x) = E[Y(1)|X = x]$$

is estimated by a base learner, such as any supervised machine learning or regression model, using the observations in the treatment group $\{X_i^1, Y_i^1\}$. Second, the response function for patients in the control arm,

$$\mu_0(x) = E[Y(0)|X = x]$$

is estimated by another base learner, using the control group observations $\{X_i^0, Y_i^0\}$. Denote the estimates for both responses by $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$, then the CATE estimate is given by

$$\hat{\theta}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x). \tag{1.5.3}$$

Note that treatment assignment is no longer a feature in the model, since patients from different treatment arms are now used in estimating two separate response functions.

1.5.2.3 X-learner

Proposed by Kunzel et al. (2018), the X-learner is provably efficient under the scenario when the number of patient in one group is much larger than that in the other group. It involves three steps.

Step 1. Estimate the response functions using any base learners (as described in T-learner).

$$\mu_1(x) = E[Y(1)|X = x]$$
$$\mu_0(x) = E[Y(0)|X = x].$$

Step 2. Impute the treatment effects for patients in the treated group based on the control-response estimate, and the treatment effects for patients in the control group based on the treatment-response estimate.

$$D_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1),$$
$$D_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0.$$

Where (Y_i^1, X_i^1) and (Y_i^0, X_i^0) denote the observed response and covariate set for individual i in the treatment or control group, respectively.

Step 3. Estimate the two treatment effect functions using any supervised learning or regression methods and obtain estimates $\hat{\tau}_1$ and $\hat{\tau}_0$.

$$\theta_1(x) = E[D^1|X = x]$$
$$\theta_0(x) = E[D^0|X = x].$$

Finally the CATE is defined as a weighted linear combination of the two treatment effects estimates.

$$\hat{\theta}(x) = g(x)\hat{\theta}_0(x) + (1 - g(x))\hat{\theta}_1(x). \tag{1.5.4}$$

Here $g(x) \in \{0, 1\}$ is a weight function and it is good to use an estimate of the propensity score for g as suggested by Kunzel et al. (2018).

The base learners for meta-learners (i.e., for each treatment arm in T- and X-learner, and for overall population in S-learner) can be any machine learning or regression models. It greatly adds flexibility to the framework and makes it generalizable to any type of outcomes. For example, if the outcome of interest is continuous, the base learners could be linear regression or regression tree. While for binary outcomes, logistic regression or random forest can be used. For time-to-event outcomes, we specifically examined the following three types of models.

1.5.3 Random survival forest (RSF) model

The random survival forest model (Ishwaran et al., 2008) is a tree-ensemble nonparametric method for survival outcomes. It grows every single tree by randomly drawing bootstrap samples from original data and further randomly selecting a subset of predictors as candidates for splitting at each node. At each node, the best split is found among all binary splits defined by the selected predictors according to a splitting rule, such as the log-rank test. Finally, the model aggregates terminal nodes across all survival trees and obtain an ensemble survival prediction. RSF has become a popular survival prediction method, and it does not assume linearity among predictors. We implement the RSF model through the R package *randomforestSRC* (Ishwaran and Kogalur, 2007).

1.5.4 Bayesian accelerated failure time model (BAFT)

The Bayesian accelerated failure time model is based on the function $\log T = f(X) + W$, where $f(X)$ is a sum of Bayesian additive regression trees (BART), and W is the residual term. Henderson et al. (2020) proposed to model the distribution of W as a location-mixture of Gaussian densities by using the centered Dirichlet process (CDP) prior, which leads to a non-parametric specification. The individual trees and bottom nodes serve as model parameters with a regularization prior to allow each tree to contribute only a small part to the overall fit and thus avoid over-fitting. Technical details can be found in Henderson

et al. (2020). We implement the BAFT model through the R package *AFTrees* on Github.

1.5.5 Cox-based deep neural network (DNN) survival model

The DNN model based on the Cox model can be expressed as $h(t|X_i) = h_0(t)e^{g(X_i;\beta)}$, where the prognostic index $g(X_i;\beta)$ is an unknown function with parameters β. Traditional Cox model assumes a simple linear relationship on the prognostic index ($g(X_i;\beta) = \beta X_i$), while the DNN model can approximate various non-linear covariate structures by estimating $g(X_i;\beta)$. Sun et al. (2020) implemented the Efron's approach in the partial likelihood to handle tied events and introduced the L_1 penalty to deal with high-dimensional covariates, with the DNN loss function presented as:

$$l(\beta;X) = \frac{1}{N_D}\sum_{j\in D}\left\{\sum_{i\in H_j} g(X_i;\beta) - \sum_{l=0}^{m_j-1}\log\left(\sum_{i\in R_j} e^{g(X_i;\beta)} - \frac{l}{m_j}\sum_{i\in H_j} e^{g(X_i;\beta)}\right)\right\},$$

where D is the set of all events with size N_D and $\{t_j\}$ is the set of unique event times; H_j is the set of subjects $\{i\}$ such that $Y_i = t_j$ and $\delta_i = 1$ and m_j is the size of H_j; and R_j is the risk set satisfying $Y_i \geq t_j$. Once $\hat{g}(X_i;\hat{\beta})$ is obtained, the predicted survival probability for subject i at time t can be computed through $\hat{S}(t|X_i) = \exp\{-\hat{H}_0(t)e^{\hat{g}(X_i;\hat{\beta})}\}$.

2.0 A Simultaneous Inference Procedure to Identify Subgroups from RCTs: Application to Analysis of AMD Progression

2.1 Introduction

There has been increasing interest in discovering personalized medicine in current pharmaceutical drug development and medical research. One aspect of personalized medicine research is to tailor existing therapies to individual patients so that each patient can get the most "suitable" treatment. Another aspect is to develop new therapies that target a subgroup of patients through modern randomized controlled trials. This research focuses on the latter aspect, which is also called "targeted" or "tailored" drug development. In such a development process, researchers are concerned with finding whether there are subgroups from an overall patient population that exhibit a differential response to the treatment. The subgroup with a significantly better response to the treatment could be identified for a tailoring strategy with appropriate labeling language and reimbursement considerations in the market. The best known example of a drug targeting a subgroup of patients is Herceptin for breast cancer patients with HER2/neu over-expression (Romond et al., 2005). More recent examples of such drugs include Xalkori for non-small cell lung cancer patients with ALK translocation (Shaw et al., 2011) and Zelboraf for skin cancer patients with BRAF mutation (Flaherty et al., 2011).

In RCTs, there is usually a treatment arm and a control arm (e.g., placebo or standard-of-care). The "relative effect" between treatment and control is referred to as "treatment efficacy". How to confidently identify subgroups that exhibit enhanced treatment efficacy (from testing a large collection of markers) is a fundamental problem in targeted drug development. Consequently, how to correctly measure this treatment efficacy is critical and can be non-trivial. It depends on the nature of the disease being treated and the clinical outcome of interest. For example, in type-II diabetes, the primary clinical outcome is the decrease in HbA1c from baseline, which is a continuous and (often) normally distributed outcome. A natural efficacy measure is the difference in the mean decrease of HbA1c, relative to the

control arm. However, in studies where the primary endpoint is time-to-event, such as time to cancer remission, or binary, such as being a responder or not, and when the patient population is a mixture of subgroups with differential treatment responses, the commonly used hazard ratio HR or odds ratio OR is not a suitable efficacy measure. Lin et al. (2019) fully discussed this issue and provided a formal definition of "logic-respecting" efficacy measure. Simply, a logic-respecting efficacy measure has to satisfy the criterion that the efficacy for the combined (mixture) group has to be between the efficacies of the subgroups. This logic-respecting property is related to "collapsibility", which is defined for measures of association in the literature (Pearl, 2000) as well as for measures of causal effect (Greenland et al., 1999; Hernan and Robins, 2020; Huitfeldt et al., 2019).

Putting it in context of measuring treatment efficacy when subgroups exist, an efficacy measure is "strictly collapsible" if each subgroup having the same efficacy implies all-comers having the same efficacy as well. Thus, being logic-respecting implies collapsibility. Note that the efficacy measures and their properties (such as "logic-respecting" or "collapsible") are defined in the parameter space (i.e., population level), and these properties are specific to the measures, not the statistical models that are used to estimate these measures.

With enormous advances in genotyping technologies, high-throughput genetic data become more increasingly collected in modern RCTs for the potential of personalized medicine development. Usually either a pre-selected panel of single nucleotide polymorphisms (SNPs) or SNPs across the whole genome are genotyped to study how patients respond differently to the therapy based on their genetic makeup. For example, in our motivating study, the Age-Related Eye Disease Study (Age-Related Eye Disease Study Research Group, 1999), which is a large multi-center RCT for an eye disease, age-related macular degeneration, DNA samples of consenting participants were collected and the genome-wide typing was performed (Fritsche et al., 2013, 2016). AMD is a polygenic and progressive neurodegenerative disease, which is a leading cause of blindness in the elderly. Patients can progress to one or both forms of late-AMD: central geographic atrophy (GA) and choroidal neovascularization (CNV). Many genetic studies have shown that the development or the progression of AMD is associated with various genetic risk factors (Fritsche et al., 2016; Seddon et al., 2007; Sun and Ding, 2019; Wei et al., 2020b; Yan et al., 2018). Specifically, in two recent genomewide asso-

ciation studies for AMD progression using the AREDS data (Sun and Ding, 2019; Yan et al., 2018), where time-to-late-AMD is the outcome, multiple variants from *ARMS2-HTRA1* and *CFH* gene regions have been discovered to be associated with AMD progression. Besides association analyses where no treatment is involved, multiple research groups also investigated whether variants from these two gene regions are associated with differential treatment responses. A recent review article by Cascella et al. (2018) summarized the controversial findings. Research groups such as Klein et al. (2008) and Seddon et al. (2016) reported that genetic variants from *CFH* and *ARMS2* regions were found to be associated with differential responses to the antioxidants plus zinc treatment. However, the AREDS investigators reported no significant associations between *CFH* and *ARMS2* regions and the nutritional supplements, when multiplicity adjustment has been taken into account (Chew et al., 2015). To fully understand the effects of those nutritional supplements on AMD progression and development and to infer whether there are genetic subgroups with enhanced treatment efficacy, a rigorous statistical procedure that can simultaneously identify and infer subgroups is required. We are specifically interested in studying the treatment effects of two types of endpoints: time-to-event endpoint and binary endpoint. The time-to-event endpoint can be expressed as the time to late AMD status, while the binary endpoint refers to the 10-year progression status of AMD patients.

2.2 Time-to-event endpoint

2.2.1 Existing methods for identification of subgroups

There is a rich literature for detecting heterogeneous treatment effects across groups for time-to-event outcomes. One simple but broadly used method is to test the treatment-by-marker interaction in the CoxPH model (Cox, 1972b). However, this method cannot provide which group to target directly, nor can it provide inference on subgroup-specific efficacy. Post-hoc analyses are usually required. Another type of approach focuses on testing the existence of a subgroup (with an enhanced treatment effect) using either a logistic-Cox model

for the response in each subgroup and the latent subgroup membership (Wu et al., 2016) or a new CoxPH model including a nonparametric component for the covariate in the control group and a subgroup-treatment-interaction effect defined by a change plane (Kang et al., 2017). There are also many tree-based methods for subgroup identification. For example, RECursive Partition (RECPAM) (Ciampi et al., 1995; Negassa et al., 2005) tends to select a split that maximizes the "difference" in Cox partial likelihood between the two child nodes. Su et al. (2008) developed interaction trees where the splitting criterion is based on the test statistic for the treatment-by-split interaction within each parent node intended for splitting. The Subgroup Identification based on Differential Effect Search (SIDES) approach invented by Lipkovich et al. (2011) focuses on a direct search for subgroups with a beneficial treatment effect utilizing recursive partitioning on each individual candidate biomarker. Loh et al. (2015) developed a framework called GUIDE (Generalized, Unbiased, Interaction Detection and Estimation), which reduces the selection bias in tree-based subgroup search methods by including an additional screening step to select the best covariates adjusted for the number of possible splits. Similar ideas were described in the conditional inference trees (Hothorn et al., 2006) and extended to a more general setting of model-based recursive partitioning (Seibold et al., 2016; Zeileis et al., 2008). The Virtual Twins method developed by Foster et al. (2011b) first predicts the individual treatment effect using random forests under the potential outcome framework and then identifies subgroups by applying classification and regression trees (CART). Another group of methods aims at finding the optimal treatment regimes, including Zhao et al. (2012), Zhang et al. (2012) and Xu et al. (2015). Instead of searching for the subgroups with beneficial treatment effect, these method tend to find the best treatment for a given patient profile. Lipkovich et al. (2017) provides a thorough review on but not limited to the above mentioned methods. Interested readers may refer to Table XV in their paper for key features of commonly used subgroup identification methods. Recently Zhang et al. (2020) proposed a nonparametric method for subgroup identification using restricted mean survival time based on maximizing a value function that directly reflects the subgroup-treatment interaction effect.

The aforementioned methods are mostly machine learning-based. Another main type of approach for identifying subgroups is through multiple testing, and our proposed method

belongs to this type. The biggest advantage of our method is that it offers "strong" multiple comparisons error rate control, controlling the expected number of confidence intervals with false coverages (regardless of what the true parameter values are). Such confidence statement is challenging for machine learning methods to match. Machine learning methods often use permutation or other resampling techniques. However, to achieve "strong" multiple comparisons control, resampling-based cross-validation needs to resample through all possible parameter configurations (all true, all but one true, all but two true, and etc), which is computationally prohibitive. It is understandable that methods such as SIDES (Lipkovich et al., 2011) resample at the so-called "complete-null" only, with all the nulls being true, resulting in "weak" multiple comparisons error rate control. In the original SIDES paper, the authors discuss such "weak" control in Section 5. In our motivating disease, AMD, the markers for subgroup identification are SNPs. With linkage disequilibrium among the SNPs, the complete-null (no association) hypothesis is statistically false (Ding et al., 2018). In such circumstances, strong control of multiple comparisons error rate is desired. In addition, none of these aforementioned methods simultaneously provides inference for treatment efficacy in both targeted group and non-targeted group, and some approaches use illogical efficacy measures. The targeted treatment development process usually involves the co-development of a drug compound and a companion diagnostic tool that identifies the suitable subgroup of patients for the drug to target. Therefore, the subgroup usually needs to be "simple" (e.g., defined by one or two biomarkers) for clinical and regulatory feasibility. In this article, we develop a multiple-testing-based approach which aims to simultaneous identify and infer "simple" subgroups with enhanced treatment efficacy defined using a logical efficacy measure.

The section is organized as follows. Section 2.2.2 introduces the logic-respecting efficacy measure for time-to-event outcomes that we choose to use and its associated properties, and with that efficacy measure how we formulate the contrasts to identify subgroups and adjust for the multiplicity. Section 2.2.3 presents simulations to show finite sample performance of the proposed method and uses realistic simulations to summarize practical rules for the use of the method. Then we apply our method on the AREDS data and present our findings in Section 2.2.4. Finally, we discuss and conclude in Section 2.2.5.

2.2.2 Proposed Methods

Motivated by the AREDS data set where the SNPs were considered as markers for subgroup identification, we focus on the setting of ordinal markers that separate the population into three groups ($M = 0, 1, 2$) to illustrate our method. Brief discussions are provided in Section 2.2.5 regarding how to generalize the method to handle markers with more categories or continuous markers.

2.2.2.1 Ratio of quantile survival times and its property

In the analysis of survival outcomes, it is known that the hazard ratio is lack of collapsibility (Ding et al., 2016) and not logic-respecting (Lin et al., 2019) when subgroups exist. Therefore other treatment efficacy measures, such as the ratio or difference of (1) survival probability at a specific time point, (2) mean (restricted) survival time, and (3) quantile survival time, are considered. Ding et al. (2016) demonstrated that the ratio or difference of mean or median survival times (between Rx and C) is logic-respecting. In this manuscript, we choose ratio of quantile survival times as our efficacy measure where quantile value can be pre-specified. The interpretation of this measure is straightforward and we will also demonstrate that it has a unique property under the accelerated failure time (AFT) model that we consider.

Assume the time-to-event data fit the following AFT model:

$$\begin{aligned} \log T = \beta_0 + \beta_1 Trt + \beta_2 I(M = 1) + \beta_3 I(M = 2) + \\ \beta_4 Trt \times I(M = 1) + \beta_5 Trt \times I(M = 2) + \boldsymbol{\beta_6 X} + \sigma W, \end{aligned} \tag{2.2.1}$$

where $Trt = 0$ (C) or 1 (Rx) is the treatment assignment, $M = 0$, 1, or 2 is the marker for subgroup testing. If the marker is a SNP, then M denotes the number of minor allele (denoted as 'a') the patient carries (0 for AA, 1 for Aa, and 2 for aa). $\boldsymbol{X}$ denotes the covariates (in addition to M and Trt) that are possibly associated with the outcome, and the residual term σW can follow various distributions (e.g., extreme value distribution, which

makes the AFT model equivalent to a Weibull model). $\boldsymbol{X}$ is also called "prognostic" factors. The model in formula (2.2.1) can also be expressed with respect to the survival function:

$$S(t|Trt, M, \boldsymbol{X}) = S_0(\exp\{-(\beta_0 + \beta_1 Trt + \beta_2 I(M=1) + \beta_3 I(M=2) + \beta_4 Trt \times I(M=1) + \beta_5 Trt \times I(M=2) + \beta_6 \boldsymbol{X})\}\, t), \tag{2.2.2}$$

where $S_0(\cdot)$ is the survival function for the exponentiated residual term $e^{\sigma W}$.

Denote by $\nu_{M,\tau}^{Trt}$ the corresponding quantile survival time at which survival probability is equal to τ in each marker-by-treatment group (τ is pre-specified), and denote by r_M^τ the quantile ratio between Rx and C in each marker group. Then by setting the survival function for each group equal to τ, the corresponding quantile survival time and their ratios can be directly calculated as follows:

$$\nu_{0,\tau}^{Rx} = S_0^{-1}(\tau)e^{\beta_0+\beta_1+\boldsymbol{\beta_6 X}}, \quad \nu_{0,\tau}^{C} = S_0^{-1}(\tau)e^{\beta_0+\boldsymbol{\beta_6 X}}, \quad r_0 = \frac{\nu_{0,\tau}^{Rx}}{\nu_{0,\tau}^{C}} = e^{\beta_1},$$

$$\nu_{1,\tau}^{Rx} = S_0^{-1}(\tau)e^{\beta_0+\beta_1+\beta_2+\beta_4+\boldsymbol{\beta_6 X}}, \quad \nu_{1,\tau}^{C} = S_0^{-1}(\tau)e^{\beta_0+\beta_2+\boldsymbol{\beta_6 X}}, \quad r_1 = \frac{\nu_{1,\tau}^{Rx}}{\nu_{1,\tau}^{C}} = e^{\beta_1+\beta_4},$$

$$\nu_{2,\tau}^{Rx} = S_0^{-1}(\tau)e^{\beta_0+\beta_1+\beta_3+\beta_5+\boldsymbol{\beta_6 X}}, \quad \nu_{2,\tau}^{C} = S_0^{-1}(\tau)e^{\beta_0+\beta_3+\boldsymbol{\beta_6 X}}, \quad r_2 = \frac{\nu_{2,\tau}^{Rx}}{\nu_{2,\tau}^{C}} = e^{\beta_1+\beta_5}.$$

It can be easily seen that the ratio in each subgroup does not depend on the quantile value τ (and thus we drop τ in the super-index of r_M). More importantly, although the quantile survival time for each group depends on the baseline prognostic factors ($\boldsymbol{\beta_6 X}$), the ratio does **not**. We name this as the *covariate-invariance* property. This property is attractive as it makes the comparison (between Rx and C) simple. Further it can be shown that this property also holds in the combined groups. For example, suppose we are interested in the ratio of quantile survival times in the mixture population of $\{M = 0, 1\}$ (denoted as r_{01}^τ). We can calculate r_{01}^τ from its definition, $r_{01}^\tau = \frac{\nu_{01,\tau}^{Rx}}{\nu_{01,\tau}^{C}}$, where $\nu_{01,\tau}^{Rx}$ and $\nu_{01,\tau}^{C}$ can be obtained by solving the following equations,

$$\begin{aligned} t = \nu_{01,\tau}^{Rx} \;&:\; p_0 S_0(e^{(-\beta_0-\beta_1-\boldsymbol{\beta_6 X})}t) + (1-p_0)S_0(e^{(-\beta_0-\beta_1-\beta_2-\beta_4-\boldsymbol{\beta_6 X})}t) = \tau, \\ t = \nu_{01,\tau}^{C} \;&:\; p_0 S_0(e^{(-\beta_0-\boldsymbol{\beta_6 X})}t) + (1-p_0)S_0(e^{(-\beta_0-\beta_2-\boldsymbol{\beta_6 X})}t) = \tau, \end{aligned} \tag{2.2.3}$$

with p_0 representing the prevalence of $M = 0$ in the combined population $\{0, 1\}$. By combining the two groups at the probability level, this calculation follows the subgroup

mixable estimation (SME) principle (Ding et al., 2016). Let $x_{01,\tau}^{Rx} = e^{-\beta_0 - \boldsymbol{\beta_6 X}} \nu_{01,\tau}^{Rx}$ and $x_{01,\tau}^{C} = e^{-\beta_0 - \boldsymbol{\beta_6 X}} \nu_{01,\tau}^{C}$, then we have $r_{01}^{\tau} = \nu_{01,\tau}^{Rx} / \nu_{01,\tau}^{C} = x_{01,\tau}^{Rx} / x_{01,\tau}^{C}$. Since the solutions for $x_{01,\tau}^{Rx}$ and $x_{01,\tau}^{C}$ from equations (2.2.3) are free of $\boldsymbol{\beta_6 X}$, we also have the covariate-invariance property for r_{01}^{τ}. Note that this property holds regardless of which specific error distribution is chosen.

2.2.2.2 Confident Effect 4 contrasts (CE4) for ratio of quantile survival times

Targeted therapy development concerns about (1) whether there exists a subgroup with enhanced treatment efficacy and (2) the treatment efficacy in both targeted and non-targeted subgroups (for appropriate drug labeling and reimbursement considerations). To answer both questions simultaneously, we propose to use contrasts to directly compare efficacy between different subgroups and combination of subgroups. The markers in our motivating example are SNPs, and for each SNP, individuals can be separated to 3 genotype groups AA, Aa, aa. The traditional SNP testing problem arises from the genome-wide association study (GWAS), which aims to identify SNPs that are associated with a specific disease (and no treatment is involved). It is a common practice to test for each SNP whether it has a dominant, recessive, or additive effect. For a given SNP, assume having minor allele "a" is harmful. Then for a dominant effect, individuals carrying at least one copy of minor allele (i.e., Aa, aa) are associated with the disease, and the risks of developing the disease from individuals carrying one copy and carrying two copies are equal. For a SNP to have a recessive effect, only individuals carrying two copies of minor allele (i.e., aa) are associated with the disease. Lastly, for a SNP to have an additive effect, the risk of developing the disease is linearly associated with the number of minor allele the individual carries.

Some commonly used methods for testing SNP association include the 2-degrees-freedom F-test, which tests the complete null ($H_0 : \mu_{aa} = \mu_{Aa} = \mu_{AA}$) against the existence of a non-zero contrast (Lettre et al., 2007); the linear trend test which tests the complete null against an additivity alternative; and the MAX3 test, where the maximum test statistics under three genetic models is used to denote the significance of a single SNP (Lettre et al., 2007). However, these methods all focus on testing the complete null against some specific

alternatives, and as demonstrated by Ding et al. (2018), the complete null can be statistically false for (almost) all SNPs as long as there is a causal SNP. Moreover, when treatments are involved, the exact "dominant", "recessive", and "additive" effects, measured by a clinical outcome, can rarely hold. Therefore, testing a complete null versus a specific alternative can be misleading since it is highly likely none of the specific alternatives or the complete null is true in reality. More discussion on this can be found in Chapter 15 in the Handbook of Multiple Comparisons (Cui et al., 2021). To fill the gap, we propose the use of the following four contrasts with survival endpoints for testing and inferring subgroups with differential treatment effects. The simultaneous confidence intervals provide direct inference on all possible SNP effects and their confidence set can be directed toward patient targeting.

$$\begin{aligned}
&\log \kappa_{(1,2):0} = \log(\frac{r_{12}}{r_0}) = \log r_{12} - \log r_0, \quad \log \kappa_{1:0} = \log(\frac{r_1}{r_0}) = \log r_1 - \log r_0, \\
&\log \kappa_{2:(0,1)} = \log(\frac{r_2}{r_{01}}) = \log r_2 - \log r_{01}, \quad \log \kappa_{2:1} = \log(\frac{r_2}{r_1}) = \log r_2 - \log r_1. \qquad (2.2.4)
\end{aligned}$$

We drop τ in the notation as τ is pre-specified. Moreover, these contrasts are built on the log scale of the efficacy measure since our previous experience demonstrates that the normality approximation seems to work better on the log scale (as compared to the original scale) (Ding et al., 2016). In fact these four contrasts are analogous to the contrasts proposed in Ding et al. (2018) where the efficacy in their case is measured by a continuous outcome. Specifically, $\log \kappa_{(1,2):0}$ tests for a "dominant" effect (of allele a), $\log \kappa_{2:(0,1)}$ tests for a "recessive" effect, and $\log \kappa_{1:0}$ and $\log \kappa_{2:1}$ test for an "additive" effect. Note that although the "superdominant" effect (e.g., the heterozygous group Aa is different from the combined two homozygous groups $\{AA, aa\}$) is rarely seen in genetics, it can also be inferred from these four contrasts, as these four contrasts can determine the complete ordering of these three genotype groups (Ding et al., 2018). From these contrasts, we are able to tell which subgroup or combination of subgroups exhibits a differential efficacy as compared to its complementary group.

Without assuming the direction of the marker effect is known (i.e., without knowing whether carrying the minor allele a is beneficial or harmful), we propose to use two-sided simultaneous confidence intervals on these four contrasts so that we can identify differential subgroup(s) and infer their efficacy simultaneously. In the situation of a confirmatory trial when one has enough prior information about the direction of the marker effect, one may

consider using four one-sided simultaneous confidence intervals. Note that level $100(1-\alpha)\%$ simultaneous confidence intervals for those contrasts effectively form a level-α interaction test: reject the null hypothesis of no interaction between Treatment effect (Trt) and marker group (M) if at least one of the confidence intervals does not contain zero. Moreover, this formulation of assessing "interaction" effect is advantageous toward patient targeting as it allows decision-making based on clinically meaningful differences (reflected from confidence intervals on efficacy comparisons) instead of a mere statistical significance (such as the p-value from an interaction test). To estimate the four contrasts from equations (2.2.4) under model (2.2.1), we propose the following three steps and name this approach as "CE4-Survival":

1. Estimate all the parameters in the survival model (e.g., parameters related to the distribution of error term and $\beta_1, \ldots, \beta_6$ in model (2.2.1)).
2. Estimate r_0, r_1, r_2, r_{01} and r_{12} and their variance covariance using estimates obtained from Step 1.
3. Calculate the four contrasts CE4 in equations (2.2.4) and obtain their joint asymptotic distribution.

The estimated variance covariance matrices in Steps 2 and 3 can be obtained using the Delta method. Note that in Step 2, the Delta method for implicitly defined random variables (Benichou and Gail, 1989) needs to be applied since the quantile survival times in the combined groups are not explicitly defined, but rather derived from solving equations in (2.2.3). Details are provided in the Appendix.

The estimated CE4 asymptotically follows a multivariate normal distribution and the simultaneous confidence intervals can be then derived as follows. We compute the quantile q such that the four simultaneous confidence intervals

$$\log(\hat{\kappa}_g) - q\hat{s}_{gg} < \log(\kappa_g) < \log(\hat{\kappa}_g) + q\hat{s}_{gg},\ g = \{(1,2){:}0,\ \ 2{:}(0,1),\ \ 1{:}0,\ \ 2{:}1\}$$

have a coverage probability $1-\alpha$, that is, the joint probability

$$Pr\Big[\frac{|\log(\hat{\kappa}_g) - \log(\kappa_g)|}{\hat{s}_{gg}} < q,\ g = \{(1,2){:}0,\ \ 2{:}(0,1),\ \ 1{:}0,\ \ 2{:}1\}\Big] = 1-\alpha,$$

where $\hat{s}^2_{gg}$ is the variance estimate for $\log(\hat{\kappa}_g)$. The *qmvnorm* function in R package {mvtnorm} can be used by inputting $1-\alpha$ and the 4-dimensional estimated correlation matrix. Meanwhile, the p-value can be obtained from the multivariate normal distribution. If any of the four contrasts does not cover 0, it suggests that there exists subgroup(s) with differential treatment efficacy.

2.2.2.3 Multiplicity adjustment across biomarkers

In targeted treatment development, typically a large collection of markers need to be tested in order to identify subgroups. Therefore, there are two families of inferences need to be considered: within a marker and across markers. Specifically, strong control of familywise error rate (FWER) for inference within a marker is desired, since the consequence of an incorrect inference may target a wrong subgroup, which is dire. The simultaneous confidence intervals obtained from our CE4-Survival method appropriately controls the within-marker FWER. While the error rate for inference across multiple markers can be controlled less stringently, since multiple candidate markers can be identified for tailoring (which may indicate largely overlapped subgroups to target), and therefore the *per family* error rate seems acceptable.

Suppose there are a total of K markers to be tested. Denote by V_k the number of confidence intervals that fail to cover the true values for the k^{th} marker. Then the FWER for the k^{th} SNP is $\alpha_k = P(V_k > 0) = E\{I_{(V_k>0)}\}$. For inference across SNPs, the *per family* error rate is $E(V_*) = E\left\{\sum_{k=1}^{K} I_{(V_k>0)}\right\} = \sum_{k=1}^{K} P(V_k > 0) = \sum_{k=1}^{K} \alpha_k$, where V_* is the number of markers with at least one of its confidence intervals failing to cover its true value. The simple *additive* adjustment proposed by Ding et al. (2018) can be applied. If the desired expected number of falsely rejected hypothesis *per family* m is chosen, then the familywise α_k for each marker is set to be $\frac{m}{K}$ for all markers. Note that this is *not* the as same as the Bonferroni probabilistic adjustment for setting $\alpha_k = \frac{\alpha}{K}$. The Bonferroni adjustment only allows α false discoveries on average, where α is the per panel FWER, usually a small number such as 0.05. While the additive adjustment allows for m false discoveries, where m is a pre-specified positive integer.

When SNPs are the biomarkers to define subgroups, the screening process seems similar to a GWAS. However, our proposal controls for *per family* error rate instead of the commonly used false discovery rate (FDR). In GWAS, it is plausible *biologically* that the vast majority of the SNPs are not associated with the specific disease. However, when treatments are involved, the biological processes become more complex, and zero-nulls of no-difference (e.g., phrased as $H_0 : \log \kappa_{(1,2):0} = 0$) are statistically false for almost all SNPs as long as there exists a causal SNP, which was first observed in the setting of Ding et al. (2018), where the treatment efficacy was simulated based on a single causal SNP with no random error being added. It was found that practically all other SNPs would appear "associated" with the outcome (as sample size reaches infinity) when analyzed in a SNP-by-SNP fashion. The reason is that most SNPs are not "orthogonal" to each other, and thus any SNP will appear somewhat associated with treatment outcome as long as the distribution for proportions of being $\{AA, Aa, aa\}$ in this SNP and the causal SNP are not independent, which is most of the cases. When there are no zero-nulls statistically, the "false" discovery seems lame, and the *per family* rate is preferred by providing more meaningful candidates. In the rest of the paper, we use an AFT-Weibull model to demonstrate the performance of the proposed method.

2.2.3 Simulation Studies

2.2.3.1 Single SNP simulations

First, we conducted simulations to investigate the finite sample performance of the proposed CE4-Survival method on analyzing one SNP with three scenarios: (1) No SNP effect, i.e., Rx is *not* efficacious for any genotype group; (2) The allele a has a dominant beneficial effect on Rx; (3) The allele a has a recessive beneficial effect on Rx. The SNP was simulated from a multinomial distribution with ($P_{aa} = 0.16, P_{Aa} = 0.48, P_{AA} = 0.36$) (corresponding to minor allele frequency (MAF) of 0.4). Survival times were first simulated from AFT model (2.2.1) with extreme value errors (equivalent to a Weibull model), where scale $\lambda = e^{-\beta_0} = 2$ and shape $k = \frac{1}{\sigma} = 1.25$. The parameters $(\beta_1, \ldots, \beta_6)$ were set to be $(0, 0.64, 0.64, 0, 0, 0)$, $(0, 0.64, 0.64, 0.48, 0.48, 0)$ and $(0, 0.64, 0.64, 0, 0.48, 0)$ for the three scenarios respectively.

The censoring times were generated from an independent uniform distribution $U(a, b)$ with a and b chosen to yield 25% and 50% censoring rates. We chose the quantile τ as 0.5 which corresponds to the median survival time. Then true values of the CE4 contrasts using the ratio of median survival as the efficacy measure for each scenario are: (1) $(1, 1, 1, 1)$, (2) $(1.62, 1.27, 1.62, 1)$, and (3) $(1.12, 1.62, 1, 1.62)$. We ran 1000 simulations with sample size 500 for each treatment arm and the results are summarized in Figure 2.2.1. Across all the scenarios, the biases of the CE4 estimates are minimal and the coverage probabilities for the simultaneous confidence intervals are all close to 95%. Larger variations are observed in biases of $\hat{\kappa}_{2:(0,1)}$ and $\hat{\kappa}_{2:1}$, especially under scenario 3. This is because under the recessive effect setting, Rx is only efficacious in $\{aa\}$ patients, which is a small proportion of the total population. Therefore, the contrasts involving the comparison between $\{aa\}$ and other group have larger variances.

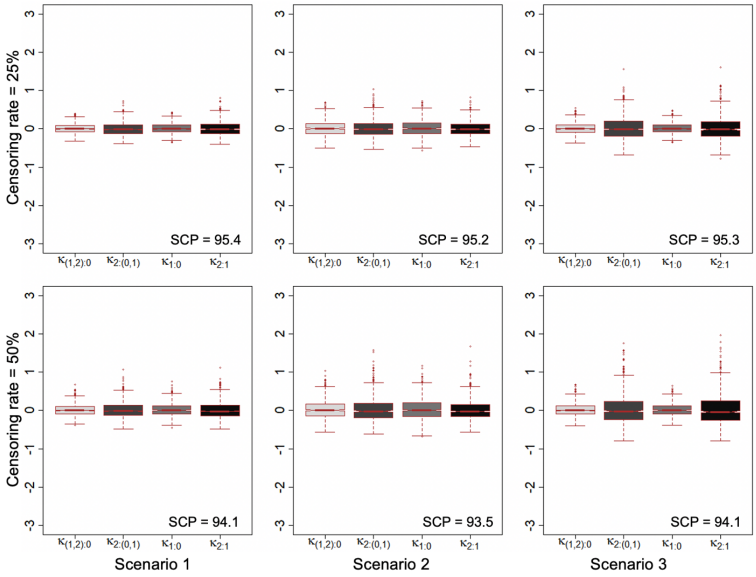

Figure 2.2.1: Finite sample performance of CE4-Survival on single SNP simulations: box plot of the biases of CE4 estimates and simultaneous coverage probability (SCP) under Weibull distribution.

To assess the robustness of CE4-Survival method under model mis-specification, we simulated survival time from two other settings where the Weibull model does not hold. In the first setting, the data were simulated from a Gompertz distribution where $S(t) =$

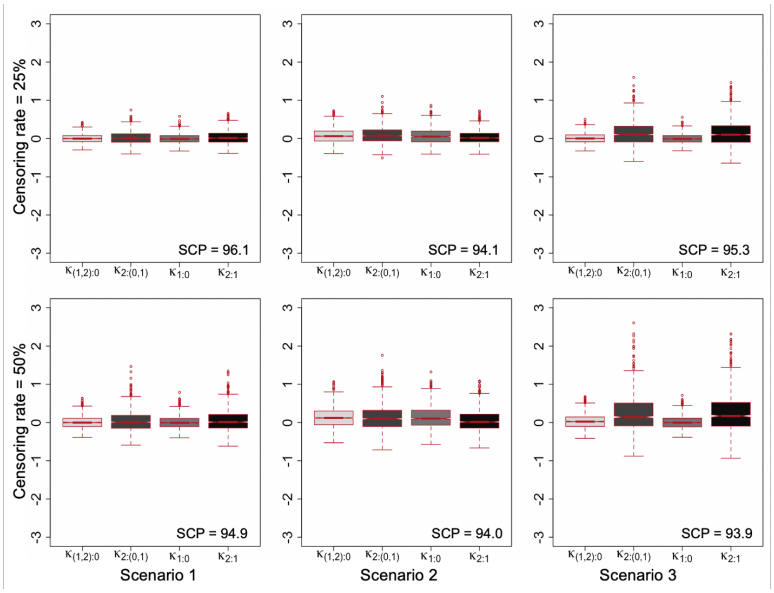

Figure 2.2.2: Finite sample performance of CE4-Survival: box plots of the biases of CE4 estimates and SCP under Gompertz distribution.

$\exp\{\frac{(1-e^{\alpha t})\lambda e^{\beta X}}{\alpha}\}$. The parameters are set to be $\lambda = 0.5$, $\alpha = 0.25$ and $(\beta_1, \ldots, \beta_6) = (0, -0.8, -0.8, 0, 0, 0)$, $(0, -0.8, -0.8, -0.6, -0.6, 0)$ and $(0, -0.8, -0.8, 0, -0.6, 0)$ to give the following true values of the CE4 contrasts: (1) $(1, 1, 1, 1)$, (2) $(1.54, 1.21, 1.54, 1)$, and (3) $(1.12, 1.54, 1, 1.54)$. Such data fit a proportional hazards model but not an AFT model. In the second setting, the data were simulated from an AFT model with error W generated from a standard logistic distribution. The model parameters (β_s, σ) were set to be the same as in the Weibull model described previously in the AFT model (2.2.1) and give the following true values of CE4 contrasts : (1) $(1, 1, 1, 1)$, (2) $(1.82, 1.30, 1.82, 1)$, and (3) $(1.09, 1.82, 1, 1.82)$. Therefore the data fit a proportional odds model but not a Weibull model. Then we applied the CE4-Survival model with Weibull distribution for each scenario. Results from 1000 simulations with sample size 500 for each treatment arm are summarized in Figure 2.2.2 and Figure 2.2.3. Although the fitted model is mis-specified, with data generated from Gompertz distribution, the comparisons between treatment effects of different subgroups (i.e., the four estimated contrasts) are relatively robust with small biases and the coverage probabilities

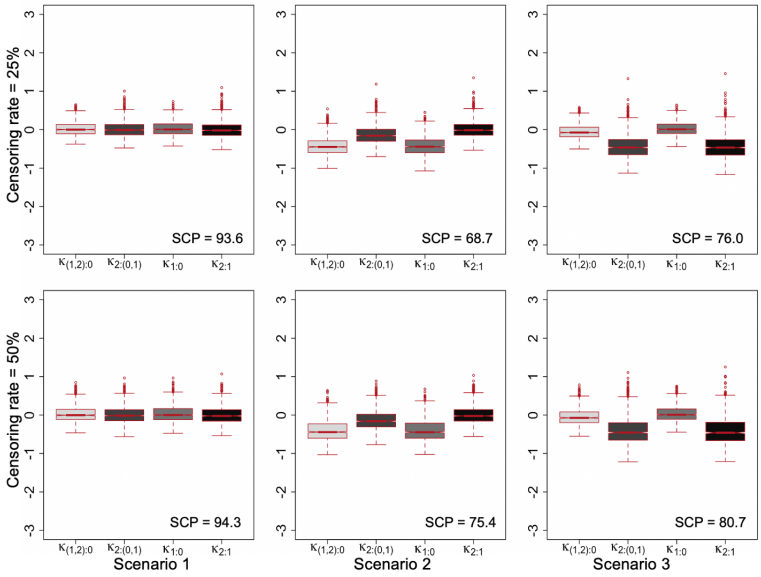

Figure 2.2.3: Finite sample performance of CE4-Survival: box plots of the biases of CE4 estimates under log-logistic distribution.

for the simultaneous confidence intervals are close to the nominal level. For data generated from the log-logistic distribution, the biases are still minimal for the no effect scenario. However, for the two scenarios where differential treatment efficacy exists, we observe biases and under-estimated coverage probabilities. Therefore, the proposed CE4-Survival method has some robustness against model mis-specification, but like any parametric-based approach, model fitting diagnostics are necessary in practice.

2.2.3.2 Chromosome-wide realistic simulations

To understand the performance of the proposed CE4-Survival method in real genetic settings with a large number of SNPs, we used the real SNP data from AREDS. Among those participants who had DNA collected and genotyped, we randomly selected 1000 Caucasian participants and "assigned" them in a 1:1 ratio to the treatment Rx and a control C. A variant rs2284665 from the well-known AMD risk gene region *ARMS2-HTRA1* on Chromosome 10 was selected as the causal SNP, and the minor allele of this variant was assumed to have

a dominant beneficial effect on Rx. We kept the three genotype groups (defined by the causal SNP) balanced between Rx and C. The progression times were simulated from the Weibull model with $\lambda = 2$ and $k = 1.25$. The $\boldsymbol{\beta}s$ were set to be $(0, 0.32, 0.32, 0.51, 0.51, 0)$, corresponding to $(\kappa_{(1,2):0}, \kappa_{2:(0,1)}, \kappa_{1:0}, \kappa_{2:1}) = (1.90, 1.51, 1.90, 1)$. The censoring rate was set to be 25%.

We analyzed chromosome 10 using our CE4-Survival model and filtered the SNPs with less than three patients in each genotype group within each treatment arm, which resulted in a total of 268,053 SNPs. We set $m = 10$, allowing on average 10 out of $\sim$ 270,000 SNPs with at least one confidence interval failing to cover its true value, which is equivalent to setting the α_K level at 3.73×10^{-5} ($= \frac{m}{K} = \frac{10}{268{,}053}$). A total of 37 SNPs were identified, among which 30 SNPs are from the *ARMS2-HTRA1* region, including the causal SNP rs2284665. Other seven SNPs belong to six different gene regions, which are distance away from the causal gene region. Figure 2.2.4A plotted the positions of these SNPs relative to the causal SNP, with y-axis ($-\log_{10}(p)$) showing the significance level of each SNP. Figure 2.2.4B plotted MaxEff vs $-\log_{10}(p)$, where MaxEff (maximal effect) is defined as the maximum absolute value among the estimated CE4 contrasts that do not cover zero. The causal SNP has the smallest p-value ($= 8.52 \times 10^{-10}$) and a MaxEff of 2.20. Note that some top SNPs have very large MaxEff values. For example, SNP rs10857454 from the *C10orf128-C10orf71-AS1* region has the largest MaxEff of 29.7, while its p-value is not very small ($= 3.21 \times 10^{-6}$, close to the threshold). We caution against the situation when a huge effect size is seen, since such a huge effect for treatment efficacy is clinically unlikely. For this specific SNP, it is not surprising to see the corresponding confidence interval for $\kappa_{2:(0,1)}$ is very wide and the effective patient population only consists 1.5% of the total population.

To further investigate the relationship between the identified SNPs and the causal SNP, we proposed a novel SNP ***cross-talk*** plot. It is based on a ternary diagram using barycentric coordinates to display the proportion of three variables that sum to one. Specifically, we projected the percentages of the AA, Aa, and aa categories of the causal SNP rs2284665 in each of these categories of a given top SNP onto the triangular diagram, and connect the points with lines. If the SNP is highly correlated with the causal SNP in terms of the distribution of AA, Aa, and aa, the percentages will be close to (1,0,0), (0,1,0) and (0,0,1),

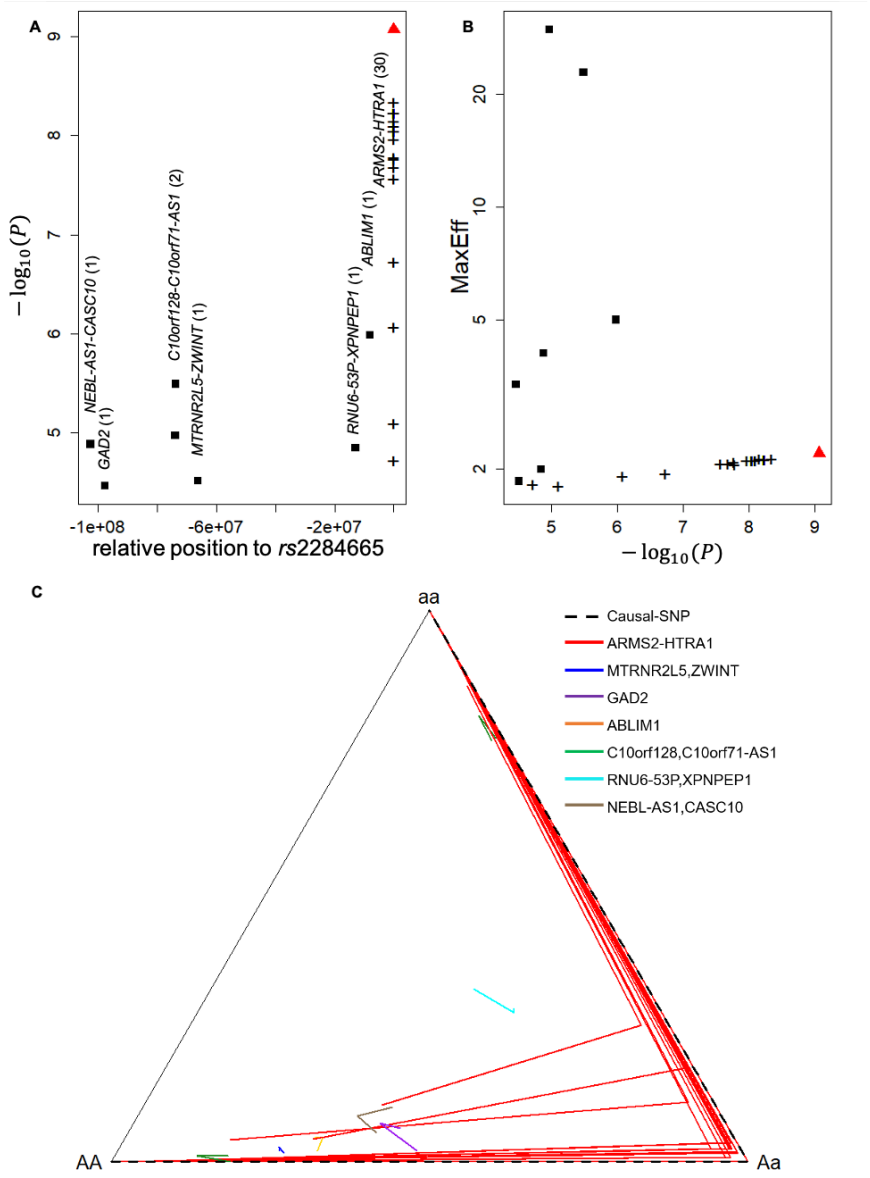

Figure 2.2.4: 37 identified SNPs from one chromosome-wide realistic simulation. **A**: -$\log_{10}(p.\text{CE4})$ *vs.* relative position to the causal SNP; **B**: the maximum effect among CE4 *vs.* -$\log_{10}(p.\text{CE4})$; the red triangle denotes the causal SNP $rs2284665$ and '+'s are the SNPs that are from the same region with the causal SNP. The rests are from other gene regions; **C**: SNP cross-talk plot.

and thus the connected line segments will be long and lie closely to the two edges of the triangle. Otherwise, the three dots will be close to each other to give a short angle. For example, in Figure 2.2.4C, the causal SNP has a perfect match in terms of the percentages with itself so the three points are the vertexes of the triangle, which makes the connected line segment coincide with the edges $AA - Aa$ and $Aa - aa$ (denoted by the dashed lines). From the plot, all 30 SNPs from *ARMS2-HTRA1* region are highly correlated with the causal SNP, indicated by the long red line segments, which explains why they have been identified by CE4-Survival. For the 7 SNPs from other regions, their line segments are all short, suggesting they might have been identified due to randomness. Note that the choice of "m" can be subjective and should be combined with prior knowledge if applicable. In this simulation study, when using $m = 5$ instead (i.e., $\alpha_k = 1.87 \times 10^{-5}$), there are 34 "significant" SNPs with at least one confidence interval that does not cover 0. Thirty of them are in the ARMS2 region, and the causal SNP is within these 30 SNPs. We recommend to use an "m" value that corresponds to an α_k of order 10^{-6} to 10^{-5} based on our previous experience with genome-wide subgroup identification using continuous or binary outcomes.

The chromosome-wide realistic simulation were repeated 100 times, where the SNP data are all the same but the progression times and censoring times are different due to randomness from the model. By setting $m = 10$, on average there are 61 SNPs identified per run with a total of 3292 SNPs being picked at least once. The causal SNP was picked 90 out of 100 times and the distribution of the ranks is shown in the stem-and-leaf plot (upper panel in Figure 2.2.5). Note that 84 out of 90 times the rank of the causal SNP was among top 30 and 52 times it was among top 10, indicating that our CE4-Survival is robust in identifying the true causal SNP. The lower panel in Figure 2.2.5 summarizes all the identified SNPs from all 100 runs in terms of their relative position to the causal SNP and their frequencies of being picked up. We found that 3256 of the 3292 SNPs (98.9%) were only picked less than 5 times out of 100 repeated simulations, which are highly likely due to randomness. While for SNPs close to the causal SNP and located in the same *ARMS2-HTRA1* gene region, the selection probability is much higher, among which 27 SNPs were identified for more than 80% of the times. From this repeated chromosome-wide simulations, we confirmed that there are possibilities that some SNPs are picked by random error but the true causal SNP and its

surrounding SNPs can be identified with high probabilities by CE4-Survival. Moreover, due to the existence of linkage disequilibrium among SNPs, it is very unlikely that an isolated SNP will be the true causal SNP.

Based on the observations from our realistic simulations, we recommend the following rules to guide the selection of "candidate" SNPs from those identified by CE4-Survival: (1) There are multiple SNPs (≥ 3 for example) being picked from the same gene region; (2) The MaxEff should not be unrealistically large; and (3) The targeted group should be a reasonable proportion (not too small or large, e.g., $5\% - 95\%$) of the total population. Note that, the numbers in the parenthesis (≥ 3 or $5\% - 95\%$) are just examples (as they are subjective), but not gold standard. After deciding the "candidate" SNP(s), the next step is to identify the targeted population based on the CE4 results. We provide a flowchart (Figure 2.2.6) to guide the subgroup identification procedure. Note that the flowchart identifies combined subgroup for targeting if both individual subgroup(s) and combined subgroup show beneficial treatment effects. For example, when both SCI1 and SCI2 are positive, $\{aa\}$ is a subgroup for the treatment to target since $\log \hat{r}_2$is significantly larger than $\log \hat{r}_{01}$. However, since SCI1 is also positive, which indicates that the combined group $\{Aa, aa\}$ is more efficacious than $\{AA\}$. In this case, we conclude the subgroup for treatment to target is $\{Aa, aa\}$. For the label consideration, if there is evidence that $\{aa\}$ is more efficacious than $\{Aa\}$ (e.g., SCI4 is positive), it is certainly important to note that in the label. Therefore, even though subgroups maybe inferred by checking some of the contrasts (not all), all contrasts are useful for inferring a full picture of the treatment efficacy.

2.2.4 Application to AREDS Data

2.2.4.1 More on AREDS

AREDS is a large multi-center RCT sponsored by the National Eye Institute to evaluate the effect of antioxidants and/or zinc on delaying the progression of AMD (Age-Related Eye Disease Study Research Group, 1999). The original study includes four treatment arms: placebo, antioxidants, zinc and the combination of antioxidants and zinc, where the last treatment then becomes the "AREDS formula" dietary supplements which are now available

rs2284665

0	11111111111111111112223333333444445555667777778899
1	001345556666888999
2	11112222223345567
3	4
4	2
7	3
9	6
13	3
14	9

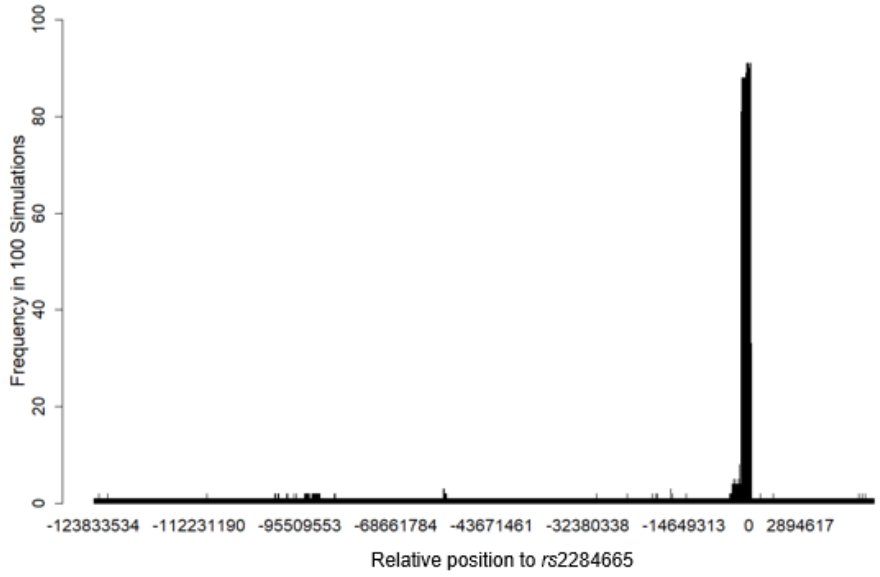

Figure 2.2.5: **Upper:** stem-and-Leaf plot for the distribution of the ranks of the Causal SNP; **Lower:** present frequency of the identified SNPs in 100 simulations.

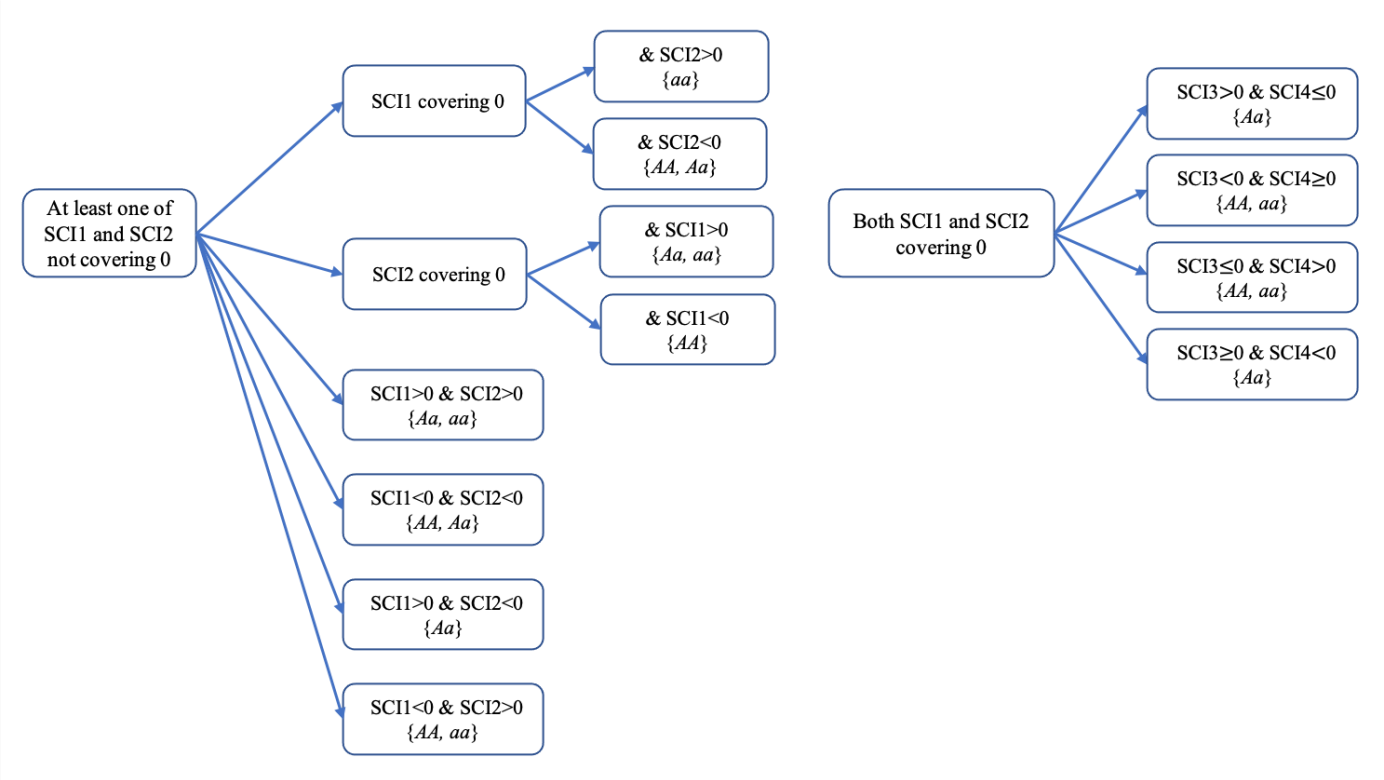

Figure 2.2.6: Flowchart of determining the targeted population based on CE4 results.

in various drug stores. However, the treatment effects of the non-placebo arms on delaying the late AMD progression are not statistically significant (Ding et al., 2017). Among the four arms, we specifically investigated participants in the placebo arm (C) and the combination of antioxidants and zinc treatment arm (Rx), which included 1,170 Caucasian participants with both eyes free of advanced AMD progression when entering the study. The outcome is the time-to-late-AMD from the first progressed eye, where late-AMD is defined as the severity score reaches 9 or above (9=GA, 10=central GA, 11=CNV, 12=central GA and CNV). As shown in Table 2.2.1, age, sex and smoking status do not differ between the two treatment arms. However, the baseline AMD severity score is significantly higher among patients who were randomized to the Rx group, as compared to patients who were randomized to placebo (4.0 ± 2.2 vs 2.6 ± 2.1). This is as expected and is due to the randomization design: patients free of AMD at baseline can only be randomized to the placebo or antioxidants arms, but

not the combination arm, and thus the baseline severity score needs to be adjusted in the analysis. The overall non-progressed rate is about 75%, so we chose $\tau = 0.75$ as the quantile of interest (to avoid extrapolation of the survival curves).

2.2.4.2 CE4-Survival on AREDS

We first evaluated the effect of "AREDS formula" on time-to-progression using a Weibull regression model, adjusting for the known risk factors including age, smoking status and baseline severity score (Chakravarthy et al., 2010; Ding et al., 2017; Group, 2001). From Section 2.2.3.1 we noticed that the model can be sensitive to the mis-specification, so we graphically checked the model fitting using Cox-Snell residuals for the "null" Weibull regression model for time-to-late-AMD on AREDS data, where the alignment of the curve with the 45 degree line indicates the overall fitting is fine. The estimated ratio of $75th$ quantile progression-free time for Rx and C is 0.91 with $p = 0.12$. It suggests that the combination of antioxidants and zinc does not seem to be effective in slowing down the disease progression in the overall population, which is consistent with previous findings (Ding et al., 2017). Then we applied CE4-Survival method using Weibull distribution to analyze all common variants (i.e., MAF ≥ 0.05) across 22 autosomal chromosomes, resulting in a total of 3,837,556 SNPs. The upper panel of Figure 2.2.7 presents the Manhattan plot of this genome-wide CE4-Survival analysis result. By setting $m = 10$, a total of 46 SNPs meet the significance threshold of $2.61 \times 10^{-6} (= m/K)$. These SNPs are from nine gene regions on seven chromosomes. Following the recommendation rule we proposed in Section 2.2.3.2, there are three gene regions each with at least four SNPs meeting the p-value threshold and they are labeled in the Manhattan plot: *CHST3-SPOCK2* on CHR 10 (4 SNPs), *ESRRB-VASH1* on CHR 14 (30 SNPs), and *C19orf44-CALR3* on CHR 19 (6 SNPs). We examined the correlation between all 46 identified SNPs using the cross-talk plot and presented the result in the middle panel of Figure 2.2.7. We picked rs147106198 (*ESRRB-VASH1* region on CHR14), which has the smallest p-value $(= 7.00 \times 10^{-8})$ as the reference SNP. It can be seen that the other 29 SNPs from the same *ESRRB-VASH1* region are highly correlated with this top SNP rs147106198, indicated by the long edges of the red segments. The other two gene regions,

CHST3-SPOCK2 and *C19orf44-CALR3* are not highly correlated with this *ESRRB-VASH1* region, although multiple SNPs within each region are highly dependent on each other (denoted by overlapped segments in green or blue). In this case, there may be more than one causal SNP that leads to the differential treatment effects. Note that the quantile τ needs to be pre-specified and it is based on prior knowledge. To examine the effect of different τ values on the subgroup identification, we conducted sensitivity analyses using $\tau = 0.50$ and $\tau = 0.25$. In summary, the results are very consistent when different τ's are used. In all three choices of τ, the same three gene regions were selected (*ESRRB-VASH1*, *CHST3-SPOCK2* and *C19orf44-CALR3*). The same set of 46 SNPs presented in the paper were found to be significantly associated with differential treatment effect in all three τ choices. When $\tau = 0.5$, these 46 SNPs were identified with slightly different ranking from the scenario with $\tau = 0.75$. When $\tau = 0.25$, a total of 53 SNPs were identified, where the additional 7 SNPs are from *C19orf44-CALR3* on chromosome 19 and *CHST3-SPOCK2* on chromosome 10. Both regions are already identified when setting $\tau = 0.75$. What's more, the top SNP with the smallest p-value (rs147106198 on *ESRRB-VASH1* region) when $\tau = 0.75$ remains to be the top one when $\tau = 0.50$ and $\tau = 0.25$. We used it as our candidate marker for further discussion.

The lower panel of Figure 2.2.7 demonstrates the treatment effect profiles and simultaneous confidence intervals for rs147106198, where the efficacy profile may suggest a dominant beneficial effect of a. The CE4 simultaneous confidence intervals confirm that the targeted group is $\{Aa, aa\}$ combined since the confidence intervals of $\log(\kappa_{(1,2):0})$ and $\log(\kappa_{1:0})$ are above the zero line. This targeted group consists about 52% of the total patients, a reasonably high proportion of the entire population. The estimated ratio of $75th$ quantile progression-free times in the targeted and non-targeted groups (between Rx and C) are presented in Table 2.3.6, which are 1.44 and 0.57 for $\{Aa, aa\}$ and $\{AA\}$, respectively, indicating that the combination of antioxidants and zinc extends the progression time for 44% compared to the placebo in the targeted group. Table 2.3.6 also provides the baseline characteristics of targeted and non-targeted population based on rs147106198, in which the patients in the targeted group do not differ from the patients in the non-targeted group. It indicates that the enhanced benefit from the treatment in the targeted population is plausibly due to the

genetic difference rather than the demographic or clinical differences.

Table 2.2.1: Baseline characteristics of the AREDS data

Number of subjects	All (n=1170)	Placebo (n=754)	Antioxidants and Zinc (n=416)	p-value*
Age				0.309
Mean (SD)	68.4 (4.9)	68.3 (4.8)	68.6 (4.9)	
Median (Range)	68.2 (55.3-81.0)	68.0 (55.3-81.0)	68.7 (55.5-79.5)	
Sex (n, %)				0.289
Female	655 (56.0)	413 (54.8)	242 (58.2)	
Male	515 (44.0)	341 (45.2)	174 (41.8)	
Smoking (n, %)				0.758
Never Smoked	571 (48.8)	371 (49.2)	200 (48.1)	
Former/Current Smoker	599 (51.2)	383 (50.8)	216 (51.9)	
Baseline AMD severity score				<0.001
Mean (SD)	3.2 (2.2)	2.6 (2.2)	4.0 (2.1)	
Median (Range)	2.0 (1.0-8.0)	1.0 (1.0-8.0)	4.0 (1.0-8.0)	
Status (n, %)				<0.001
Progressed	269 (23.0)	133 (17.6)	136 (32.7)	

*p-value is based on two-sample t test or Pearson Chi-square test for continuous or categorical variables

It should be noted that based on different SNPs, the suggested targeted population may vary. As shown in the middle panel of Figure 2.2.7, SNPs from the same *ESRRB-VASH1* region are highly correlated with the top SNP $rs147106198$. If we chose another SNP, $rs77000175$ in the region to be the candidate marker, the targeted population is about 50% of the total population, which overlaps with the targeted population indicated by $rs147106198$ by 94.7%. In another example, if a top SNP from *CHST3-SPOCK2* on CHR 10 is considered as the candidate marker (e.g., rs1245576), the targeted population is about 65.8% of the total population, which overlaps only 67.1% of the targeted population determined by the top SNP. Hence, our CE4-Survival method provides reliable and interpretable candidate targeted populations for consideration, while the final decision on which population to target

involves many other considerations such as development of companion diagnostics, labeling, marketing, and reimbursement.

2.2.4.3 Investigation on two reported gene regions using AREDS data

To help elucidate the controversial findings regarding whether genetic polymorphisms of *CFH* and *ARMS2-HTRA1* alter the treatment efficacy of AREDS formula, we closely checked six SNPs from these two regions and their results are presented in Table 2.2.3. Note that rs412852, rs1061170, and rs3766405 from *CFH* and rs10490924 from *ARMS2-HTRA1* have been previously investigated (Assel et al., 2018; Seddon et al., 2016; Vavvas et al., 2018). We also examined the SNPs with the smallest CE4-based p-value from each region, which are rs7522681 and rs11200647. None of these SNPs meets the significance threshold of 2.61×10^{-6}, although three SNPs from *CFH* region meet the nominal level of 0.05. We further investigated rs412852 from *CFH* and it seems our CE4-Weibull result suggests the combination group $\{AA, Aa\}$ exhibit better treatment efficacy compared to its complementary group $\{aa\}$, which is similar to the findings from Seddon et al. (2016) and Assel et al. (2018). However, it is worthwhile to note that from our genomewide CE4 analysis, none of these SNPs ranked top (Table 2.2.3). Therefore, with appropriate multiplicity adjustment, neither *CFH* nor *ARMS2-HTRA1* region has SNPs showing significant association with treatment efficacy, which is consistent with the conclusion indicated by Chew et al. (2015).

2.2.4.4 Validation on AREDS2

AREDS2 was another independent large multi-center RCT of AMD (Chew et al., 2012). It was designed to evaluate the effect of refined AREDS formulations on AMD progression, as compared to the original AREDS formula. Participants of AREDS2 were more severe at baseline and the follow-up time was only about half of the AREDS's follow-up time. There were four arms with AREDS supplements being the control arm (all other three arms are AREDS supplements plus additional nutrients). Since there is no placebo arm in AREDS2, we cannot apply CE4-Survival to identify subgroups with enhanced efficacy of AREDS supplements. Instead, we investigated the patient's response to the same AREDS supplements

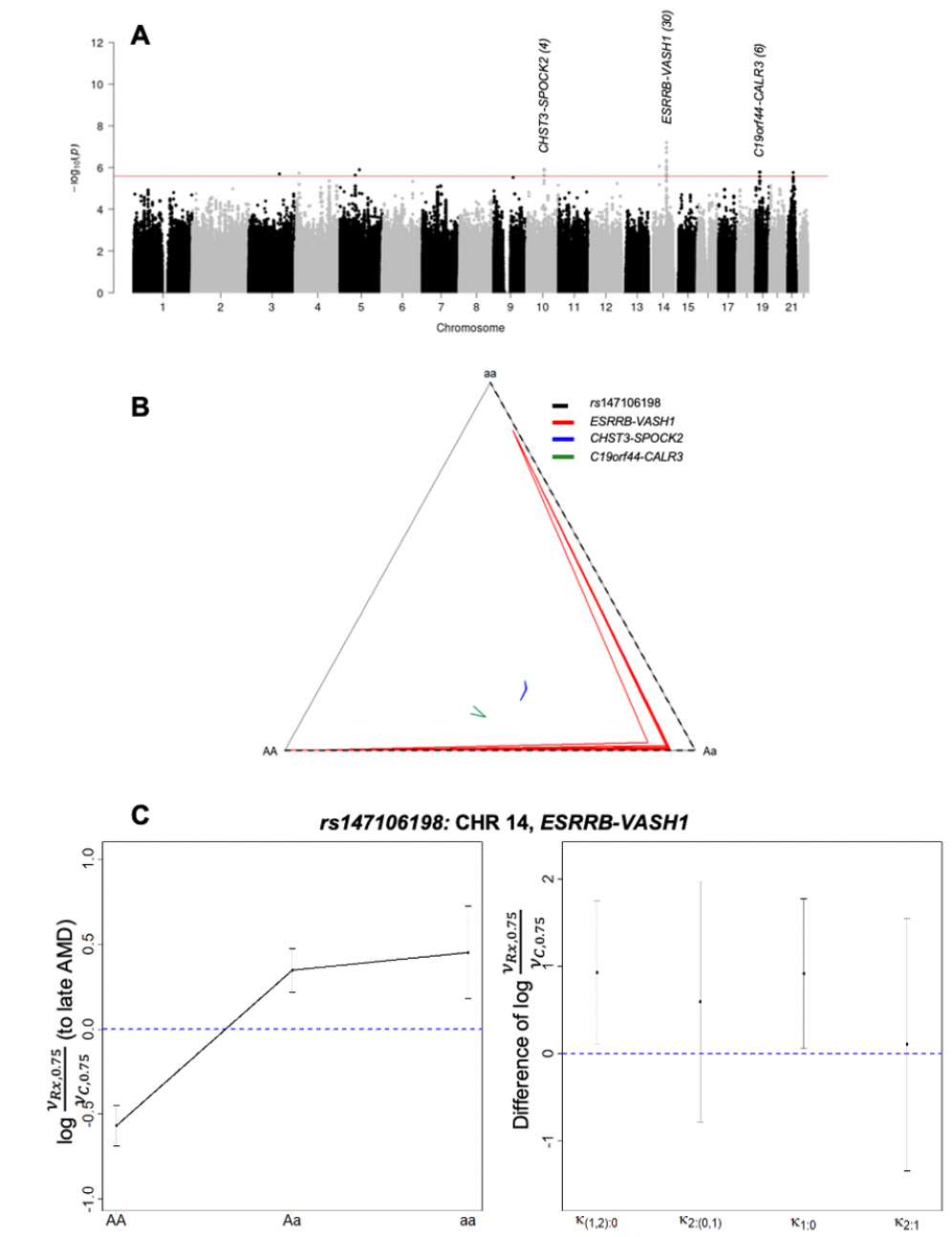

Figure 2.2.7: **A:** Manhattan plot from the genomewide CE4-Survival analysis on AREDS data; **B:** SNP cross-talk plot for 40 identified SNPs in relationship with the most top SNP rs147106198; **C:** Treatment effects and CE4 estimates for the top SNP rs147106198 (Left: treatment profile using log of the ratio of $75th$ quantile survivals; Right: CE4 estimates and simultaneous confidence intervals).

Table 2.2.2: Characteristics of targeted and non-targeted populations

	*rs*147106198: chr14, *ESRRB-VASH1* region		
	Targeted	Non-targeted	p-value
# of subjects (n,%)	605 (51.7)	565 (48.3)	
Treatment efficacy $\frac{\hat{\nu}_{Rx,0.75}}{\hat{\nu}_{C,0.75}}$ † (SE)	1.44 (1.01)	0.57 (1.01)	$6.99 \times 10^{-8\star}$
Age			0.560
Mean (SD)	68.5 (4.8)	68.4 (4.9)	
Median (range)	68.2 (55.3-81.0)	68.2 (55.8-80.5)	
Sex (n, %)			0.982
Female	338 (55.9)	317 (56.1)	
Male	267 (44.1)	248 (43.9)	
Smoking (n, %)			0.169
Never Smoked	283 (46.8)	288 (51.0)	
Former/Current Smoker	322 (53.2)	277 (49.0)	
Treatment (n, %)			0.510
Placebo	384 (63.5)	370 (65.5)	
Antioxidant + Zinc	221 (36.5)	195 (34.5)	
Baseline AMD severity score			0.487
Mean (SD)	3.1 (2.2)	3.2 (2.2)	
Median (range)	2.0 (1.0-8.0)	3.0 (1.0-8.0)	

†: $\hat{\nu}$ denotes the estimated quantile progression time

$\star$: p-value is from the corresponding CE4 contrast when simultaneous type I error is controlled, without adjusting for cross-SNP multiplicity

arm to check whether we observe similar differential response patterns between the targeted and non-targeted groups (identified from AREDS) in AREDS2. Table 2.2.4 presents the patient characteristics within the targeted and non-targeted groups (determined by SNP rs147106198), separately for AREDS and AREDS2. None of these baseline risk factors differs between targeted and non-targeted populations in each study. The only difference is the genotype in terms of rs147106198. Figure 2.2.9 compares the progression-free Kaplan-Meier

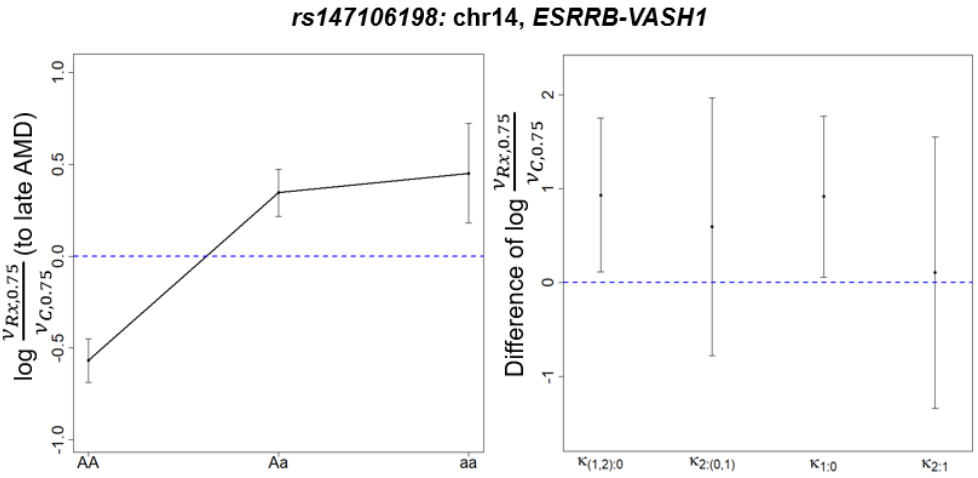

Figure 2.2.8: Top identified SNP from AREDS, $rs147106198$. **Left:** treatment profile using log of the ratio of $75th$ quantile survivals; **Right:** CE4 results by taking the difference between the log ratio of $75th$ quantile progression time to late AMD in the presented contrasts.

Table 2.2.3: CE4 results of six selected SNPs from *CFH* and *ARMS2-HTRA1* regions

Gene	SNP	p.CE4	rank.CE4
CFH	$rs7522681$	3.74×10^{-4}	3843
	$rs412852$	9.49×10^{-4}	9021
	$rs1061170$	1.22×10^{-2}	71651
	$rs3766405$	0.38	1614412
ARMS2-HTRA1	$rs11200647$	7.75×10^{-2}	378198
	$rs10490924$	0.77	3059694

curves between the targeted and non-targeted groups within each study. In both studies, the targeted population shows an obvious better progression-free profile than the non-targeted population (the log-rank test $p = 0.00011$ and 0.013 respectively). Therefore, we successfully validated our identified targeted group based on rs147106198 in AREDS2.

Table 2.2.4: Characteristics of targeted and non-targeted populations in AREDS and AREDS2

	AREDS			AREDS2		
	Targeted	Non-targeted	p-value*	Targeted	Non-targeted	p-value*
# of subjects (n,%)	221 (53.1)	195 (46.9)		164 (51.1)	157 (48.9)	
Age			0.696			0.242
Mean (SD)	68.6 (4.8)	68.8 (5.0)		70.2 (7.4)	71.1 (7.8)	
Median (range)	68.4 (55.5-78.5)	68.9 (56.1-79.5)		71.0 (51.0-85.0)	71.0 (53.0-86.0)	
Sex (n, %)			0.700			0.824
Female	131 (59.3)	111 (56.9)		96 (58.5)	89 (56.7)	
Male	90 (40.7)	84 (43.1)		68 (41.5)	68 (43.3)	
Smoking (n, %)			0.806			0.587
Never Smoked	108 (48.9)	92 (47.2)		77 (47.0)	68 (43.3)	
Former/Current Smoker	113 (51.1)	103 (52.8)		87 (53.0)	89 (56.7)	
Baseline AMD severity score			0.375			0.347
Mean (SD)	4.0 (2.1)	4.2 (2.1)		6.5 (1.1)	6.6 (1.0)	
Median (range)	4.0 (1.0-8.0)	4.0 (1.0-8.0)		7.0 (2.0-8.0)	7.0 (2.0-8.0)	
Genetic risk score			0.157			0.840
Mean (SD)	0.99 (0.13))	1.01 (0.14)		1.09 (0.13)	1.09 (0.13)	
Median (range)	0.99 (0.62-1.30)	1.01 (0.71-1.34)		1.10 (0.71-1.37)	1.10 (0.58-1.40)	

*p-value is based on two-sample t test or Pearson Chi-square test for continuous or categorical variables

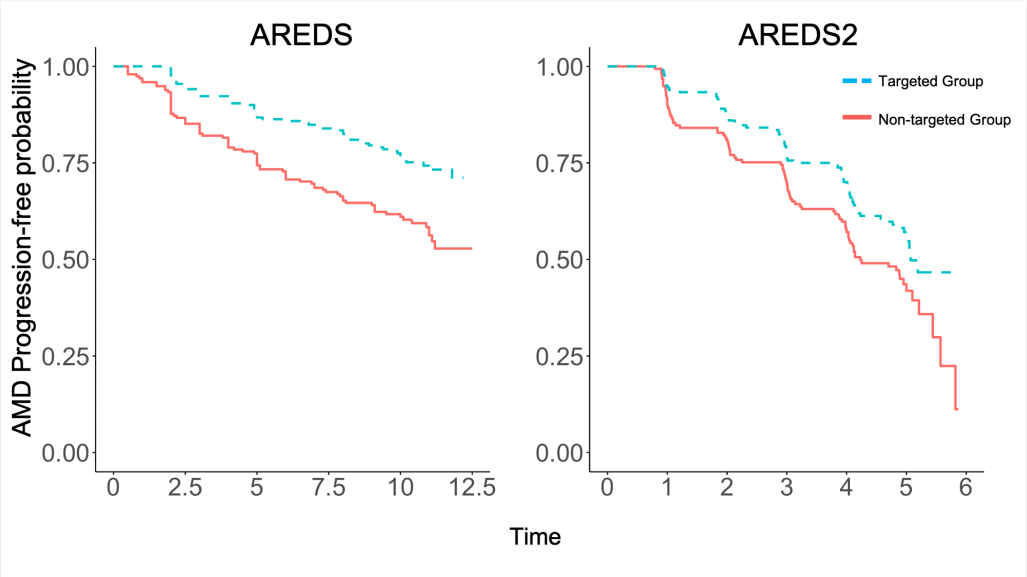

Figure 2.2.9: Kaplan-Meier curves for targeted/non-targeted patients taking AREDS supplements in AREDS and AREDS2, where the subgroup is defined by $rs147106198$.

2.2.5 Conclusion and Discussion

In this work, We develop and implement a new method to confidently identify and infer subgroups in modern randomized clinical trials with time-to-event outcomes. Different from machine learning based approaches, our CE4-Survival, derived from the fundamental multiple testing principle, provides simultaneous confidence intervals on contrasts that directly compare the treatment efficacy between subgroups or combination of subgroups. The contrasts are built upon a logic-respecting efficacy measure, the ratio of quantile survival times, which is easy to interpret and enjoys the unique covariate invariance property. Our CE4-Survival adjusts for multiplicity both within and across the markers. It rigorously combines two error rate controls, family-wise error rate control within marker, and *per family* error rate control across markers. Such error control is appropriate for drug development, as it allows flexibility in the exploration of multiple candidate markers, while being confident in the subgroup to target from any selected marker.

Our realistic simulation studies used the SNP data taken from individuals who participated in the AREDS, where the allele frequency and linkage disequilibrium are preserved. Therefore, these studies can provide recommendations of practical rules for identifying "candidate" markers. Finally, we successfully applied our method on AREDS data to identify subgroups that exhibit enhanced treatment efficacy with combination of antioxidant supplements and zinc in delaying AMD progression. We further validated the subgroups defined by the top SNP rs147106198 from *ESRRB-VASH1* region using the data from an independent AREDS2 study. SanGiovanni et al. (2013) found that the estrogen related receptor beta (*ESRRB*) was associated with CNV AMD. Wakusawa et al. (2008) first demonstrated the angiogenesis modulation of *VASH* is involved in the pathological process of AMD. Later Zeng et al. (2012) inferred that AREDS supplements is likely to affect both angiogenesis and endothelial-macrophage interactions. Thus our findings provide new perspectives on the differential treatment efficacy for AMD.

Although we use the SNP testing to demonstrate our method, the key elements of the method apply to broader scenarios with all kinds of markers. When the marker separates the patient population into more groups, additional contrasts need to be considered to obtain the complete ordering of the treatment efficacy. When the marker is with three groups but categorical instead (not ordinal as SNPs are), one may consider using a modified CE5 contrasts to build inference on comparing the combined $\{0, 2\}$ group and $\{1\}$ by adding the contrast $\log \kappa_{(0,2):1} = \log r_{02} - \log r_1$. Similar steps described in Section 2.2.2.2 can be applied. The current version of our method only handles discrete markers and more work is required to generalize it for continuous markers. In doing that, one may borrow the idea from Liu et al. (2016) which considers all candidate thresholds for a continuous marker when deriving simultaneous confidence intervals. Note that, unlike machine learning based methods, our CE4 method is not designed to test multiple markers simultaneously in a joint model.

AFT model is used to build CE4 contrasts in the current method, since it directly links to the logic-respecting treatment efficacy measure we use: the ratio of quantile survival ratio. In addition, it shows a good covariate-invariance property that makes interpretation of the treatment effect free of other prognostic factors. However, it does not mean the CE4 can only couple with the AFT model. Actually, it can be generalized to any survival model

framework as long as the treatment efficacy in the mixed population is carefully derived following the SME principle. For example, one may consider using Cox models. In order to derive the asymptotic distribution of CE4 contrasts, the variance-covariance matrices of the quantile survival time in each group and the combined group need to be constructed, which requires the variance-covariance matrix of the inverse function of Breslow estimates of baseline cumulative hazard.

In this paper, we did not account for the variability of the estimates of p_0 and p_1 (i.e., we did not treat them as parameters). Here, p_0 denotes the prevalence of $M = 0$ in the combined group $\{0, 1\}$, and p_1 denotes the prevalence of $M = 1$ in the combined group $\{1, 2\}$. They are used in calculating the quantile survival time in $\{0, 1\}$ group and $\{1, 2\}$ group. This is in concordance with the "OM" (observed margins) option in the LSmeans statement of SAS (Proc GLM, Proc Mixed), where the empirical estimate of prevalence is used, and it is not treated as a parameter in the inference procedure. One can also use the "population level" prevalence (if SNP is the marker, its prevalence can be derived from the minor allele frequency in a public genetic database, TOPMED, for example). However, the variability of the estimates of p_0 and p_1 can be accounted in our CE4 method (if one choose so). The key modification is on the parameter set in the Delta method for implicit random variables (which is described in Section 5 of Supplemental Materials). We explored this option in our simulation studies and the results are very similar to the situation without accounting for the variability of the estimates of p_0 and p_1.

2.3 Binary endpoints

2.3.1 Relative risk with log-linear model

With binary outcomes, a logistic regression model is often fitted and the OR is commonly used to measure treatment efficacy. However, it is not suitable when there exits a mixture of populations, which is demonstrated in Lin et al. (2019). Specifically, the OR in the combined group is not guaranteed to be bounded within the ORs in the two separate groups. That

is, OR_{01} (true value, not estimated) can be outside of [OR_0,OR_1], where the indices '0', '1', and '01' denote the individual or combined group. This is partly because OR_{01} can not be expressed as a weighted combination of OR_0 and OR_1, i.e., OR is not (strictly) collapsible. Both Lin et al. (2019) and Huitfeldt et al. (2019) gave similar counter examples.

Another efficacy measure that is frequently used for binary outcomes is RR. In our motivating example, it is defined as the ratio of the probability of progression to late-AMD in 10 years between those who took the treatment and those who took the placebo. It has been mathematically proven that the RR is logic-respecting with $RR_0 \leq RR_{01} \leq RR_1$ always holds. Note that using whether OR or RR as the efficacy measure reflects a choice of a clinically meaningful summary statistic to characterize the drug effect, and it is independent of what statistical model is used to fit the data. In this manuscript, we demonstrate that using RR as efficacy measure under a log-linear model has a unique and nice property for subgroup identification determined by SNPs.

Assume the binary data fit the following model:

$$g\{Pr(Y=1|Trt,M,X)\} = \beta_0 + \beta_1 I(Trt=1) + \beta_2 I(M=1) + \beta_3 I(M=2) + \beta_4 I(Trt=1)\times I(M=1) + \beta_5 I(Trt=1)\times I(M=2) + \boldsymbol{\beta_6 X}, \quad (2.3.1)$$

where $Y=1$ indicates the event of interest, g is the link function, $Trt=0$ (C) or 1 (Rx) is the treatment assignment, and $M=0$, 1, or 2 denote the genotype group AA, Aa, or aa. $\boldsymbol{X}$ represents all prognostic factors that are regardless of treatment assignment. Then the relative risk (RR) for each genotype group can be derived as follows:

$$\begin{aligned}
RR_0 &= \frac{Pr(Y=1|Trt=1,M=0,\boldsymbol{X})}{Pr(Y=1|Trt=0,M=0,\boldsymbol{X})} = \frac{g^{-1}(\beta_0+\beta_1+\boldsymbol{\beta_6 X})}{g^{-1}(\beta_0+\boldsymbol{\beta_6 X})},\\
RR_1 &= \frac{Pr(Y=1|Trt=1,M=1,\boldsymbol{X})}{Pr(Y=1|Trt=0,M=1,\boldsymbol{X})} = \frac{g^{-1}(\beta_0+\beta_1+\beta_2+\beta_4+\boldsymbol{\beta_6 X})}{g^{-1}(\beta_0+\beta_2+\boldsymbol{\beta_6 X})},\\
RR_2 &= \frac{Pr(Y=1|Trt=1,M=2,\boldsymbol{X})}{Pr(Y=1|Trt=0,M=2,\boldsymbol{X})} = \frac{g^{-1}(\beta_0+\beta_1+\beta_3+\beta_5+\boldsymbol{\beta_6 X})}{g^{-1}(\beta_0+\beta_3+\boldsymbol{\beta_6 X})}.
\end{aligned}$$

The commonly used link functions for binary outcomes include logit link, log link and probit link. When log link is applied, the RRs for each genotype group can be simplified as

$$RR_0 = e^{\beta_1},\ RR_1 = e^{\beta_1+\beta_4},\ \text{and}\ RR_2 = e^{\beta_1+\beta_5}.$$

It is worthwhile mentioning that although the probability of $Y = 1$ for each genotype group depends on the prognostic factors ($\boldsymbol{\beta_6 X}$), the ratios are all free of these covariates. This is the "covariate-invariance" property we mentioned in Section 2.2.2.1. This property makes the comparison between treatment arms straightforward and simple. However, if a logistic regression is applied (i.e., g is the logit link), this property will not hold any more. For example, the RR for the AA group under the logit link is:

$$RR_0 = \frac{1/\{1 + e^{-(\beta_0+\beta_1+\boldsymbol{\beta_6 X})}\}}{1/\{1 + e^{-(\beta_0+\boldsymbol{\beta_6 X})}\}} = \frac{1 + e^{-(\beta_0+\boldsymbol{\beta_6 X})}}{1 + e^{-(\beta_0+\beta_1+\boldsymbol{\beta_6 X})}},$$

which contains $\boldsymbol{\beta_6 X}$. Therefore, we recommend to considering the log-linear model.

Besides that the RRs of individual groups from the log-linear model are covariate-invariant, it can be shown that this property also holds in the combined groups. Following the SME principle (Ding et al., 2016), denote by π_0 the prevalence of $M = 0$ in the combined population $\{0, 1\}$, then $1 - \pi_0$ is the prevalence of $M = 1$ in the combined population. The probability of $Y = 1$ within Rx and C in the $\{0, 1\}$ combined population can be expressed as:

$$p_{01}^{Rx}(Y = 1) = \pi_0 p_0^{Rx}(Y = 1)+(1-\pi_0)p_1^{Rx}(Y = 1) = \pi_0 e^{\beta_0+\beta_1+\boldsymbol{\beta_6 X}}+(1-\pi_0)e^{\beta_0+\beta_1+\beta_2+\beta_4+\boldsymbol{\beta_6 X}},$$

$$p_{01}^{C}(Y = 1) = \pi_0 p_0^{C}(Y = 1) + (1 - \pi_0)p_1^{C}(Y = 1) = \pi_0 e^{\beta_0+\boldsymbol{\beta_6 X}} + (1 - \pi_0)e^{\beta_0+\beta_2+\boldsymbol{\beta_6 X}}.$$

Then the RR for the combined group can be calculated as:

$$RR_{01} = \frac{p_{01}^{Rx}(Y = 1)}{p_{01}^{C}(Y = 1)} = e^{\beta_1}\frac{\pi_0 + (1 - \pi_0)e^{\beta_2+\beta_4}}{\pi_0 + (1 - \pi_0)e^{\beta_2}}.$$

Similarly, one can derive that $RR_{12} = \frac{\pi_1 e^{\beta_2+\beta_4}+(1-\pi_1)e^{\beta_3+\beta_5}}{\pi_1 e^{\beta_2}+(1-\pi_1)e^{\beta_3}}$. It can be seen that these RRs are also free of $\boldsymbol{\beta_6 X}$ and thus the covariate-invariant property holds for the combined groups as well.

2.3.2 CE4 for relative risks

Inspired by CE4-Survival method, we propose the following four contrasts to get a complete ordering of the treatment efficacy in all subgroups and their combinations using RRs.

$$\begin{aligned} &\log \kappa_{(1,2):0} = \log RR_{12} - \log RR_0,\ \log \kappa_{1:0} = \log RR_1 - \log RR_0, \\ &\log \kappa_{2:(0,1)} = \log RR_2 - \log RR_{01},\ \log \kappa_{2:1} = \log RR_2 - \log RR_1. \end{aligned} \tag{2.3.2}$$

Since we focus on binary outcomes in this article and we name the proposed method as "CE4-Binary". We propose to use four two-sided simultaneous confidence intervals on the four contrasts in (2.3.2). It has been proved that this approach would guarantee to control the family-wise error rate (FWER) strongly from testing four contrasts simultaneously (see Theorem 4 of Hsu and Berger (1999)). As indicated by Ding et al. (2018), these two-sided simultaneous confidence intervals are equivalent to testing the following eight one-sided null hypotheses where each one is to test an inequality against its complement (i.e., H_0: $\log \kappa_{(1,2):0} \leq 0$ vs H_a: $\log \kappa_{(1,2):0} > 0$) rather than testing a zero null (such as H_0 : $\log \kappa_{(1,2):0} = 0$).

$$\begin{aligned} &H^{\leq}_{(1,2):0} : \log \kappa_{(1,2):0} \leq 0, \quad H^{\leq}_{(0,1):2} : \log \kappa_{(0,1):2} \leq 0, \quad H^{\leq}_{1:0} : \log \kappa_{1:0} \leq 0, \quad H^{\leq}_{1:2} : \log \kappa_{1:2} \leq 0, \\ &H^{\leq}_{2:(0,1)} : \log \kappa_{2:(0,1)} \leq 0, \quad H^{\leq}_{0:(1,2)} : \log \kappa_{0:(1,2)} \leq 0, \quad H^{\leq}_{2:1} : \log \kappa_{2:1} \leq 0, \quad H^{\leq}_{0:1} : \log \kappa_{0:1} \leq 0. \end{aligned}$$

We refer to Ding et al. (2018) for detailed reasons why testing zero-null hypotheses is not appropriate in this setting. With these four simultaneous confidence intervals, one can tell which allele is beneficial and the possible effect size of each effect. For example, if the lower bound of the confidence interval of contrast $\log \kappa_{(1,2):0}$ in (2.3.2) is greater than zero (assuming a larger value implies a better efficacy), then it indicates that the minor allele a is beneficial and the effect could be dominant.

With binary outcomes, the contrast estimates can be obtained from fitting the log-linear model (2.3.1), where each RR (for individual genotype groups and the two combined groups)

has been derived in Section 2.3.1. On log scale, the four contrasts are:

$$\begin{aligned}
\log \kappa_{(1,2):0} &= \log\{\pi_1 e^{\beta_2+\beta_4} + (1-\pi_1)e^{\beta_3+\beta_5}\} - \log\{\pi_1 e^{\beta_2} + (1-\pi_1)e^{\beta_3}\}, \\
\log \kappa_{2:(0,1)} &= \beta_5 - \log\{\pi_0 + (1-\pi_0)e^{\beta_2+\beta_4}\} + \log\{\pi_0 + (1-\pi_0)e^{\beta_2}\}, \\
\log \kappa_{1:0} &= \beta_4, \\
\log \kappa_{2:1} &= \beta_5 - \beta_4. \qquad (2.3.3)
\end{aligned}$$

Let $A_1 = \pi_1 e^{\beta_2+\beta_4} + (1-\pi_1)e^{\beta_3+\beta_5}$, $A_2 = \pi_1 e^{\beta_2} + (1-\pi_1)e^{\beta_3}$, $A_3 = \pi_0 + (1-\pi_0)e^{\beta_2+\beta_4}$, and $A_4 = \pi_0 + (1-\pi_0)e^{\beta_2}$. The first order partial derivative matrix of the CE4 contrasts (2.3.3) with respect to $\boldsymbol{\beta}$ can be derived as:

$$D_{CE_4}(\beta) = \begin{bmatrix}
\frac{\pi_1 e^{\beta_2+\beta_4}}{A_1} - \frac{\pi_1 e^{\beta_2}}{A_2} & \frac{(1-\pi_1)e^{\beta_3+\beta_5}}{A_1} - \frac{(1-\pi_1)e^{\beta_3}}{A_2} & \frac{\pi_1 e^{\beta_2+\beta_4}}{A_1} & \frac{(1-\pi_1)e^{\beta_3+\beta_5}}{A_1} \\
-\frac{(1-\pi_0)e^{\beta_2+\beta_4}}{A_3} + \frac{(1-\pi_0)e^{\beta_2}}{A_4} & 0 & -\frac{(1-\pi_0)e^{\beta_2+\beta_4}}{A_3} & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & -1 & 1
\end{bmatrix}$$

Denote $\Omega(\beta)$ to be the variance-covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$, which can be estimated from fitting the log-linear model in (2.3.1). The estimated variance-covariance matrix of (2.3.3) can be obtained using the Delta method: $\hat{\Sigma}_{CE_4} = \hat{D}_{CE_4}(\hat{\beta})\hat{\Omega}(\hat{\beta})\hat{D}^{-1}_{CE_4}(\hat{\beta})$.

The estimated CE4 asymptotically follows a multivariate normal distribution and the simultaneous confidence intervals can then be derived as described in 2.2.2.2 using R package `mvtnorm` (Genz et al., 2020) (details can be found from the R code in `https://github.com/yingding99/CE4-Binary`). One big advantage of the proposed CE4 method is that it can provide inference on treatment efficacy for the targeted and non-targeted groups through simultaneous confidence intervals, instead of just producing p-values.

Same procedure described in Section 2.2.2.3 applies to control the FWER within each SNP and the *per family* error rate across multiple SNPs, since it allows flexibility in the exploration of multiple candidate SNPs and it is possible that different SNPs lead to similar patient population for targeting, especially for SNPs within the tight linkage disequilibrium (LD) region.

2.3.3 Extension to bivariate binary outcomes

So far the proposed CE4-Binary method with RR being the efficacy measure is for univariate binary outcomes. In our motivating example, since AMD is a bilateral eye disease, the correlation between the two eyes of the same individual needs to be considered. Given the SNP effects on treatment efficacy is more appropriate to be interpreted on a marginal population-level than on a (conditional) subject-level, we choose to use the generalized estimating equations (GEE) approach (Liang and Zeger, 1986).

To estimate the RRs directly for binary data, usually either a log-binomial or a Poisson model (with log link) is recommended. However, model convergence issue seems to be more likely for the log-binomial model (Williamson et al., 2013). Zou (2003) proposed to use a Poisson model with the robust sandwich variance estimate. Yelland et al. (2011) and Chen et al. (2018) conducted thorough simulations to compare the performance between log-binomial and robust Poisson models for estimating RRs and concluded that the robust Poisson model is preferred when model is misspecified or predictors are continuous. Therefore in the following simulations and real-data analysis, we use the GEE-based Poisson regression model to estimate the RRs using the R package `geepack` (Halekoh et al., 2006). Assuming the conditional distribution of Y_{ij} given the predictor variables follows a Poisson distribution, with the mean response related to the predictors by the link function $\log(\mu_{ij}) = x_{ij}^T\beta$. We use $i = 1, \ldots, n$ to denote subject and $j = 1, 2$ to denote each eye of a subject. Let $Y_i = (Y_{i1}, Y_{i2})$ and $\mu_i = (\mu_{i1}, \mu_{i2})$. A Poisson-GEE model with log link solves the following generalized estimating equation for β:

$$S(\beta) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \beta})^T V_i^{-1} (Y_i - \mu_i) = 0,$$

where V_i is the working variance matrix that can be decomposed as $V_i = \phi A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$. A_i is the diagonal matrix of the 2×2 conditional variance-covariance matrix for subject i, where under Poisson distribution, the two elements are μ_{i1} and μ_{i2}. R_i is a 2×2 working correlation matrix (for example, the exchangeable structure is used in the following sections), and ϕ is a scale parameter that will be estimated by the data. Under certain regularity conditions,

the solution $\hat{\beta}$ has an asymptotic normal distribution $\hat{\beta} \overset{n\to\infty}{\longrightarrow} N(\beta, \Sigma)$, where $\Sigma = I_0^{-1} I_1 I_0^{-1}$, with $I_0 = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta}^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$ and $I_1 = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta}^T V_i^{-1} (Y_i - \mu_i(\beta))(Y_i - \mu_i(\beta))^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$.

2.3.4 Simulation Studies

2.3.4.1 Single SNP simulations

In this section, we examine the performance of CE4-Binary for testing a single SNP. Since our motivating example has bivariate binary outcomes (progression status of both eyes), we performed all of our simulations in the context of bivariate binary outcomes. Specifically, we used the R package `SimCorMultRes` to simulate bivariate binary outcome under the marginal model specification $P(Y_{ij} = 1|x_{ij}) = F(x_{ij}^T \beta)$ (Touloumis, 2019). In particular, $Y_{ij} = 1$ if $U_{ij} \leq 2x_{ij}^T \beta$, where $U_{ij} = x_{ij}^T \beta + e_{ij}$, and $e_{ij} \sim F$ for all i and j. For each subject i, the correlation structure among the clustered binary responses $\{Y_{ij} : j = 1, 2\}$ depends on $\{e_{ij} : j = 1, 2\}$. We denote r the correlation between e_{i1} and e_{i2}. Here, we chose F to be the cumulative density function of the standard logistic distribution. In `SimCorMultRes`, the authors employ the tetra-variate extreme value distribution to simulate correlated error terms (Gumbel, 2012) and details can be found in their package description.

To proceed, we simulated a single SNP, a binary treatment (Rx or C) and a bivariate binary outcome. Specifically, the SNP was simulated from a multinomial distribution with minor allele frequency (MAF) of 0.4, corresponding to probabilities $P_{aa} = 0.16$, $P_{Aa} = 0.48$, and $P_{AA} = 0.36$. We considered three efficacy scenarios: (1) no SNP effect, i.e., Rx is not efficacious for any SNP genotype group; (2) the allele a has a dominant beneficial effect on Rx; (3) the allele a has a recessive beneficial effect on Rx. Here we use $Y = 1$ to indicate occurrence of a undesirable event, for example, AMD progression. We also considered two sets of marginal outcome probabilities for each treatment and genotype group combination: (i) high probability of event across all combinations, denoted as P_{high}, and (ii) low probability of event across all combinations, denoted as P_{low}. The overall design is shown in Table 2.3.1, which presents the (marginal) event probabilities for each treatment-by-genotype combination, as well as the RR between Rx and C. To further assess the effect of correlation (between bivariate outcomes), we simulated the bivariate binary outcomes with

Table 2.3.1: Overall design of single SNP simulations. Probability in each cell refers to the marginal probability of event in each corresponding treatment-by-genotype group

	P_{high}			P_{low}		
	AA	Aa	aa	AA	Aa	aa
No effect						
Rx	0.4	0.4	0.4	0.2	0.2	0.2
C	0.6	0.6	0.6	0.4	0.4	0.4
RR	0.67	0.67	0.67	0.5	0.5	0.5
Dominant effect						
Rx	0.6	0.4	0.4	0.4	0.2	0.2
C	0.6	0.6	0.6	0.4	0.4	0.4
RR	1	0.67	0.67	1	0.5	0.5
Recessive effect						
Rx	0.6	0.6	0.4	0.4	0.5	0.2
C	0.6	0.6	0.6	0.4	0.4	0.4
RR	1	1	0.67	1	1	0.5

two different correlations, $r = 0.3$ or 0.6, using the approach we mentioned above.

We ran 1000 simulations for each scenario. The sample size is set as $n = 1000$ pairs and the treatment arm (Rx or C) is randomly assigned with a 1:1 ratio. The results of biases for the four contrast estimates (on the log scale) and the coverage probability for simultaneous 95% confidence intervals are summarized in Table 2.3.2. It can be seen that the biases are close to 0 and the empirical coverage probabilities are close to the nominal level under all scenarios. The standard deviation of 1000 estimated biases from the high probability scenario is smaller than that from the low probability scenario for all four contrasts across all settings. When correlation increases, the standard deviation of biases increases for all contrasts through all scenarios. This is because the "effective" sample size decreases as

correlation increases. In addition, we also computed the frequency that the confidence interval for $\log \kappa_{(1,2):0}$ does not cover 0 under the dominant effect setting and the confidence interval for $\log \kappa_{2:(0,1)}$ does not cover 0 under the recessive effect setting, respectively. In general, we found that the frequency increases with larger effect size (i.e., RR is more away from 1) and decreases with larger correlation, which are expected (details not shown).

We further conducted two additional sets of simulations by varying sample sizes to $n =$ 500 pairs and $n = 2000$ pairs. The results are summarized in Table 2.3.3 and Table 2.3.4. In general, in addition to the findings described above, we observed that the standard deviation of 1000 estimated biases gets smaller with increasing sample size, and we also noticed that the empirical SCPs are close to the nominal level under all studied scenarios even when $n = 500$ pairs.

2.3.4.2 Chromosome-wide realistic simulations

In this section, we further use the chromosome-wide data from AREDS to evaluate the performance of CE4-Binary in a real setting for testing a large number of SNPs. First, we randomly selected 1000 Caucasian participants who had their DNA collected and genotyped, and randomly "assigned" them to one of the treatment arms (Rx vs C) in a 1:1 ratio. We chose a common variant $rs2284665$ from the well-known AMD risk gene region $ARMS2$ on chromosome 10 as the causal SNP, and assumed the minor allele of this SNP has a dominant beneficial effect on Rx. To be specific, we assumed the marginal probability of AMD progression for the control group is 0.3 among all genotype groups, and the relative risks of AMD progression between the treatment group and the control group are (RR_{AA}, RR_{Aa}, RR_{aa}) = (1, 0.5, 0.5) for AA, Aa and aa genotype groups, respectively. The Pearson's correlation between the bivariate outcome is set as 0.3, which is estimated from the real study.

We applied our proposed CE4-Binary method to test all common SNPs (i.e., SNPs with at least 5 observations in each treatment-by-genotype group) on chromosome 10, and repeated the simulation 100 times. The average total number of SNPs being included in the analysis is 199,071. We set $m = 10$ to allow on average 10 out of $\sim$ 200,000 SNPs to have at least one confidence interval failing to cover its true value, which is equivalent to set α_K

Table 2.3.2: Mean (SD) of estimated biases for log CE4 estimates ($\log \kappa_{(1,2):0}$, $\log \kappa_{2:(0,1)}$, $\log \kappa_{1:0}$, $\log \kappa_{2:1}$) and simultaneous coverage probability (SCP) for the 95% simultaneous confidence interval (N=1000)

	$r = 0.3$		$r = 0.6$	
	P_{high}	P_{low}	P_{high}	P_{low}
	No effect			
$\log \kappa_{(1,2):0}$	0.0015 (0.106)	0.0011 (0.169)	0.0038 (0.111)	0.0008 (0.184)
$\log \kappa_{2:(0,1)}$	-0.0042 (0.135)	-0.0053 (0.229)	-0.0003 (0.158)	-0.0100 (0.248)
$\log \kappa_{1:0}$	0.0018 (0.115)	-0.0007 (0.175)	0.0031 (0.120)	-0.0003 (0.196)
$\log \kappa_{2:1}$	-0.0041 (0.147)	-0.0033 (0.237)	-0.0008 (0.170)	-0.0079 (0.263)
SCP	0.948	0.951	0.953	0.942
	Dominant effect			
$\log \kappa_{(1,2):0}$	0.0007 (0.115)	-0.0095 (0.192)	-0.0050 (0.131)	0.0123 (0.203)
$\log \kappa_{2:(0,1)}$	0.0010 (0.139)	-0.0114 (0.223)	-0.0012 (0.156)	-0.0010 (0.243)
$\log \kappa_{1:0}$	-0.0006 (0.121)	-0.0098 (0.203)	-0.0064 (0.137)	0.0105 (0.212)
$\log \kappa_{2:1}$	0.0017 (0.145)	-0.0072 (0.236)	0.0019 (0.163)	-0.0027 (0.252)
SCP	0.960	0.938	0.944	0.955
	Recessive effect			
$\log \kappa_{(1,2):0}$	0.0045 (0.122)	-0.0002 (0.195)	-0.0033 (0.138)	-0.0004 (0.220)
$\log \kappa_{2:(0,1)}$	0.0025 (0.142)	-0.0054 (0.226)	-0.0178 (0.165)	-0.0069 (0.249)
$\log \kappa_{1:0}$	0.0040 (0.130)	0.0005 (0.216)	0.0004 (0.150)	-0.0008 (0.237)
$\log \kappa_{2:1}$	0.0010 (0.152)	-0.0054 (0.248)	-0.0179 (0.180)	-0.0064 (0.267)
SCP	0.953	0.948	0.931	0.955

at approximately $5 \times 10^{-5} (\sim= \frac{10}{199,071})$ level.

Among all 100 repeated simulations, on average, there are 22 SNPs being identified by CE4-Binary per simulation run, with a total of 1082 unique SNPs being picked at least once.

Table 2.3.3: Mean (SD) of estimated biases for log CE4 estimates ($\log \kappa_{(1,2):0}$, $\log \kappa_{2:(0,1)}$, $\log \kappa_{1:0}$, $\log \kappa_{2:1}$) and simultaneous coverage probability (SCP) for the 95% simultaneous confidence interval (N=500)

	$r = 0.3$		$r = 0.6$	
	P_{high}	P_{low}	P_{high}	P_{low}
	No effect			
$\log \kappa_{(1,2):0}$	-0.0021 (0.151)	0.0037 (0.241)	-0.0065 (0.170)	0.0092 (0.254)
$\log \kappa_{2:(0,1)}$	-0.0043 (0.201)	-0.0186 (0.333)	-0.0116 (0.216)	-0.0392 (0.363)
$\log \kappa_{1:0}$	-0.0033 (0.160)	0.0026 (0.258)	-0.0066 (0.178)	0.0135 (0.271)
$\log \kappa_{2:1}$	-0.0018 (0.212)	-0.0159 (0.355)	-0.0075 (0.225)	-0.041 (0.385)
SCP	0.956	0.945	0.950	0.949
	Dominant effect			
$\log \kappa_{(1,2):0}$	0.0043 (0.168)	-0.0147 (0.275)	0.0047 (0.186)	-0.0053 (0.289)
$\log \kappa_{2:(0,1)}$	-0.0021 (0.203)	-0.0306 (0.323)	-0.0032 (0.215)	-0.0270 (0.361)
$\log \kappa_{1:0}$	0.0033 (0.178)	-0.0135 (0.286)	0.0042 (0.194)	-0.0051 (0.305)
$\log \kappa_{2:1}$	-0.0032 (0.213)	-0.0239 (0.335)	-0.0046 (0.225)	-0.0249 (0.381)
SCP	0.947	0.962	0.951	0.960
	Recessive effect			
$\log \kappa_{(1,2):0}$	-0.0051 (0.178)	-0.0003 (0.292)	-0.0105 (0.186)	0.0019 (0.323)
$\log \kappa_{2:(0,1)}$	0.0021 (0.203)	-0.0184 (0.336)	-0.0127 (0.225)	-0.0298 (0.376)
$\log \kappa_{1:0}$	-0.0077 (0.193)	-0.0018 (0.316)	-0.0105 (0.199)	0.0061 (0.352)
$\log \kappa_{2:1}$	0.0053 (0.220)	-0.0176 (0.360)	-0.0085 (0.240)	-0.0330 (0.404)
SCP	0.948	0.952	0.949	0.945

The causal SNP was identified 45 out of 100 times (and it is the most frequently identified SNP), and 29 out of the 45 times it was ranked top 30, which demonstrates the validity and robustness of our proposed CE4-Binary method. The relative positions of all identified

Table 2.3.4: Mean (SD) of estimated biases for log CE4 estimates ($\log \kappa_{(1,2):0}$, $\log \kappa_{2:(0,1)}$, $\log \kappa_{1:0}$, $\log \kappa_{2:1}$) and simultaneous coverage probability (SCP) for the 95% simultaneous confidence interval (N=2000)

	$r = 0.3$		$r = 0.6$	
	P_{high}	P_{low}	P_{high}	P_{low}
	No effect			
$\log \kappa_{(1,2):0}$	-0.0013 (0.075)	0.0009 (0.120)	0.0014 (0.083)	0.0018 (0.131)
$\log \kappa_{2:(0,1)}$	0.0001 (0.100)	-0.0082 (0.165)	0.0019 (0.106)	-0.0109 (0.171)
$\log \kappa_{1:0}$	-0.0020 (0.080)	0.0016 (0.126)	0.0004 (0.088)	0.0033 (0.138)
$\log \kappa_{2:1}$	0.0012 (0.106)	-0.008 (0.174)	0.0021 (0.114)	-0.0114 (0.179)
SCP	0.953	0.948	0.949	0.949
	Dominant effect			
$\log \kappa_{(1,2):0}$	0.0001 (0.082)	-0.0050 (0.136)	-0.0050 (0.088)	-0.0024 (0.141)
$\log \kappa_{2:(0,1)}$	-0.0003 (0.098)	-0.0035 (0.157)	0.0035 (0.101)	-0.0030 (0.173)
$\log \kappa_{1:0}$	-0.0003 (0.086)	-0.0063 (0.141)	-0.0063 (0.092)	-0.0036 (0.148)
$\log \kappa_{2:1}$	-0.0002 (0.103)	0.0062 (0.163)	0.0062 (0.107)	-0.0011 (0.182)
SCP	0.957	0.951	0.959	0.950
	Recessive effect			
$\log \kappa_{(1,2):0}$	0.0006 (0.088)	0.0050 (0.139)	0.0009 (0.094)	0.0030 (0.156)
$\log \kappa_{2:(0,1)}$	-0.0058 (0.098)	0.0011 (0.165)	-0.0018 (0.113)	0.0080 (0.175)
$\log \kappa_{1:0}$	0.0023 (0.096)	0.0037 (0.148)	0.0010 (0.100)	-0.0006 (0.168)
$\log \kappa_{2:1}$	-0.0068 (0.107)	-0.0004 (0.176)	-0.0023 (0.119)	0.0083 (0.188)
SCP	0.945	0.950	0.948	0.949

SNPs to the causal SNP and their frequencies of being identified are summarized in Figure 2.3.1. We found that 1051 out of 1082 SNPs (97.1%) were picked less than 5 times, which is highly likely due to randomness. The remaining 30 SNPs (in addition to the causal SNP)

are highly correlated to the causal SNP, with a mean Δ^2 (a correlation measure of LD, see Hill and Robertson (1968)) of 0.894 and a median Δ^2 of 0.948. We computed Δ^2 using the R package `genetics` in this study (Warnes, 2019).

The distribution of Δ^2 (with the causal SNP) for the 1051 SNPs (picked less than 5 times) and the 30 SNPs (picked greater than or equal to 5 times) is summarized in Figure 2.3.2. Clearly, the SNPs picked < 5 times have a much lower Δ^2 value as compared to the SNPs picked for ≥ 5 times. In addition, for those 30 SNPs, each of them was picked for 37 times on average, and they are all located on the same gene region (*PLEKHA1/ARMS2/HTRA1*). Finally, out of 100 simulations, 56 times there are at least one SNP from this gene region being picked. From these results we conclude: (1) the true causal SNP and its surrounding SNPs can be identified with high probabilities by our proposed CE4-Binary; (2) some noise SNPs might be identified by random chance; (3) due to the existence of LD among SNPs, the causal SNP is more likely to be among a cluster of SNPs.

We further conducted another set of realistic simulations with a different parameter setting, where the effect size is bigger, with $(RR_{AA}, RR_{Aa}, RR_{aa}) = (1, 0.4, 0.4)$ (i.e., larger differences in RRs between genotype groups). Similar results were observed and the relative positions of all identified SNPs to the causal SNP and their frequencies are summarized in Figure 2.3.3. In general, due to the larger effect size, the likelihood of identifying causal SNP or its surrounding SNPs increases. For instance, under this circumstance, the causal SNP was identified 88 out of 100 times.

2.3.5 SNP Effects on Treatment Efficacy of AREDS Data

2.3.5.1 AREDS data description

In this study, we chose the 10-year progression (to late-AMD) status as the outcome, where late-AMD is defined as the severity score reaches 9 or above (9=geographic atrophy (GA), 10=central GA, 11=choroidal neovascularization (CNV), 12=central GA and CNV). This is consistent with the AREDS report for evaluating long-term effects of AREDS treatments on AMD progression (Chew et al., 2013). We analyzed a total of 1,127 Caucasian participants from the combination treatment arm and the placebo arm who were free of

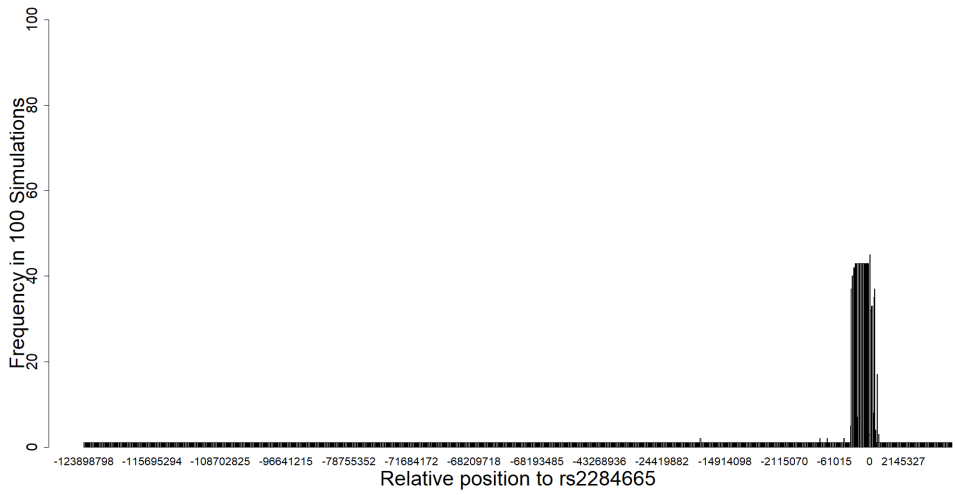

Figure 2.3.1: Histogram for the frequency of SNPs being identified in 100 chromosome-wide realistic simulations.

late-AMD in at least one eye at enrollment and did not censor from the study by 10 years. This results in a total of 2,044 eyes for analysis. DNA samples were collected and genotyped through a customized HumanCoreExome array by Illumina (Fritsche et al., 2013, 2016).

Table 2.3.5 summarizes the characteristics of our study cohort. Participants receiving placebo or AREDS formula do not differ in terms of age, sex and smoking status. However, the baseline AMD severity score is higher in the treatment arm, as compared to the placebo arm (3.9 ± 2.2 vs 2.7 ± 2.2), and so is the 10-year progression rate (32.1% vs 18.8%). This is due to the stratified randomization strategy where the least severe participants (denoted as "AMD category 1") were not randomized to receive the treatments with zinc because of its potential side effect. All other participants (AMD category 2-4) were equally randomized to any of the four treatment arms. The distribution of patients in the AMD categories confirms that participants in category 1 did not receive the antioxidants plus zinc treatment.

We then further investigated the 10-year progression rate for all 2,044 study eyes, stratified by their own and their fellow eyes' baseline (BL) severity scale. As shown in Figure 2.3.4, the progression rate increases as the study eye's own BL severity level increases (3.6%

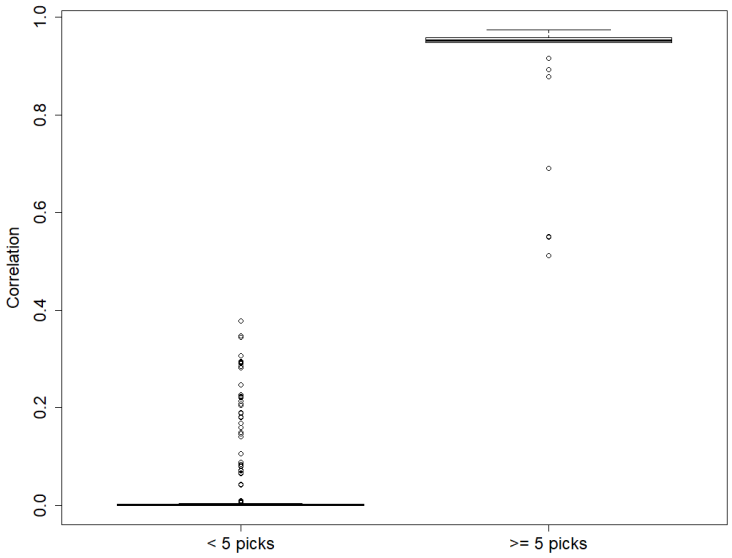

Figure 2.3.2: Distribution of correlations with causal SNP (measured by Δ^2) for SNPs picked less than 5 times and SNPs picked greater than or equal to 5 times.

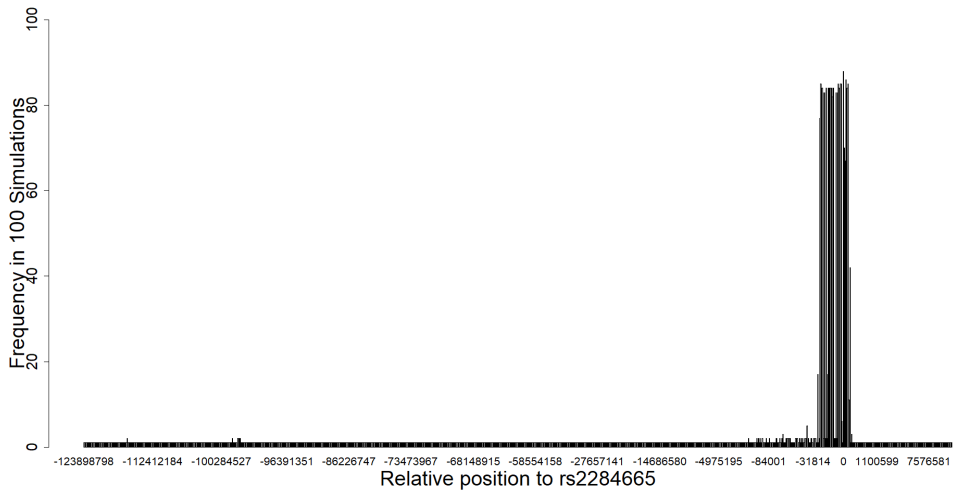

Figure 2.3.3: Histogram for the frequency of SNPs being identified in 100 chromosome-wide realistic simulations for the setting (RR_{AA}, RR_{Aa}, RR_{aa}) = (1, 0.4, 0.4).

for severity in 1-3, 40.8% for severity in 4-6, and 80.6% for severity in 7-8), indicated by the white bars. Within each category of the study eye's own BL severity level, the progression rate increases as the fellow eye's BL severity increases. For example, the 10-year progression rate for eyes with BL severity score in 4-6 increases from 18.0% to 37.9%, 61.4% and 79.4%, as the fellow eye's BL severity score increases from 1-3 to 4-6, 7-8, and 9+, respectively. As a bilateral disease, the progression of one eye highly depends on the severity of the other eye, and thus the correlation between two eyes cannot be ignored.

2.3.5.2 CE4 analysis of AREDS data

We first studied the overall treatment effect of AREDS formula on reducing the 10-year progression rate. To account for the between-eye correlation, the GEE model with Poisson distribution and log-link function was applied, adjusting for the known risk factors including age, smoking status, and baseline severity score. The estimated relative risk between Rx and C is 1.10 with $p = 0.15$, which indicates that the combination of antioxidants and zinc

Table 2.3.5: Characteristics of the AREDS data (Bivariate Binary Outcomes)

Number of subjects	All ($n = 1127$)	Placebo ($n = 677$)	Antioxidants and Zinc ($n = 450$)	p-value*
Age				0.441
Mean (SD)	68.6 (4.9)	68.5 (4.8)	68.8 (5.0)	
Median (Range)	68.3 (55.3-81.0)	68.1 (55.3-81.0)	68.7 (55.5-79.8)	
Sex (n, %)				0.159
Female	616 (54.7)	358 (52.9)	258 (57.3)	
Male	511 (45.3)	319 (47.1)	192 (42.7)	
Smoking (n, %)				0.969
Never Smoked	543 (48.2)	327 (48.3)	216 (43.0)	
Former/Current Smoker	584 (51.8)	350 (51.7)	234 (52.0)	
AREDS AMD categories (n, %)				<0.001
1	277 (24.6)	277 (40.9)	0 (0)	
2	241 (21.4)	119 (17.6)	122 (27.1)	
3	415 (36.8)	190 (28.1)	225 (50.0)	
4	194 (17.2)	91 (13.4)	103 (22.9)	
Eye-level variables				
Number of eyes	$n = 2044$	$n = 1255$	$n = 789$	
Baseline AREDS AMD severity score				<0.001
Mean (SD)	3.1 (2.3)	2.7 (2.3)	3.9 (2.3)	
Median (Range)	2.0 (1.0-8.0)	1.0 (1.0-8.0)	4.0 (1.0-8.0)	
10-year progression (n, %)				<0.001
Progressed	489 (23.9)	236 (18.8)	253 (32.1)	

$^\star p$-value is based on two-sample t-test or Chi-square test for continuous or categorical variables.

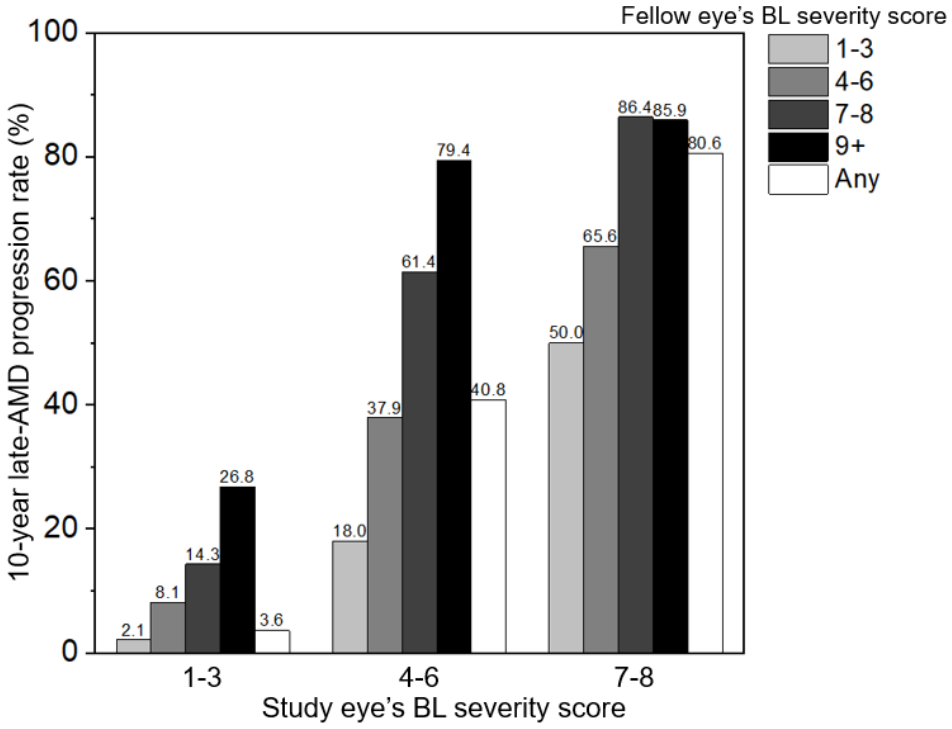

Figure 2.3.4: Progression rates (from enrollment up to 10 years) by baseline (BL) AMD severity score, stratified by the fellow eye's BL AMD severity score.

is not efficacious in the overall population, as compared to the placebo. To investigate the heterogeneity of the efficacy in the subgroups defined by a SNP, we apply our CE4-Binary method to common variants across the genome. Similar to the chromosome-wide simulation study, we kept SNPs with ≥ 5 observations in each treatment-by-genotype group, resulting in a total of $K = 3,895,495$ SNPs for analysis. The same set of risk factors was included as covariates in the model.

The Manhattan plot in Figure 2.3.5 shows the findings of the genome-wide CE4 analysis. We set $m = 10$, which corresponds to a significance level at 2.57×10^{-6} in this case. A total of 28 SNPs from seven gene regions on seven chromosomes are identified. Among those gene regions, three of them have more than 3 SNPs meeting the p-value threshold and they are labeled in the Manhattan plot: *ANGPT2-MCPH1* on chromosome 8 (5 SNPs), *CHST3-SPOCK2* on chromosome 10 (11 SNPs), and *PPM1H* on chromosome 12 (4 SNPs). The *CHST3-SPOCK2* region was also found to have multiple SNPs with heterogeneous treatment efficacy (in terms of the AREDS formula) for AMD progression in a survival analysis by Wei et al. (2020a) and is discussed in Section 2.2.4.2.

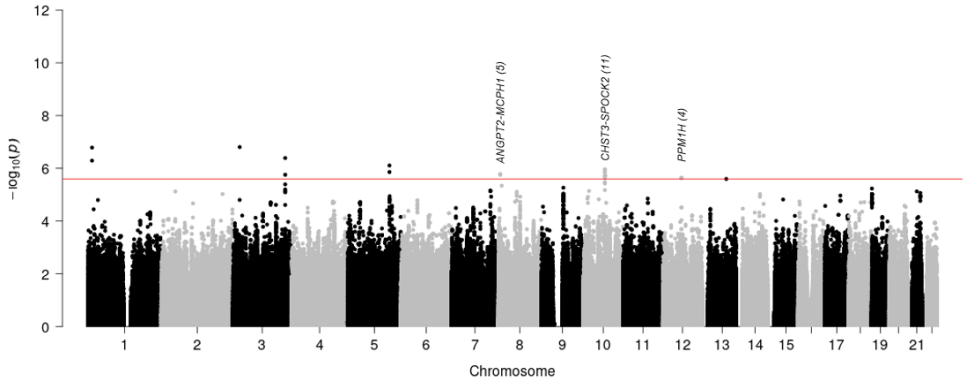

Figure 2.3.5: Genome-wide SNP effects on treatment efficacy.

Within each gene region, the identified SNPs are highly correlated in terms of Δ^2 (as shown by the heatmap in Figure 2.3.6). Therefore, we examined the subgroups determined by the top SNP (i.e., SNP with the smallest p-value) in each region. The subgroups determined by other SNPs in the same region are similar. For example, Figure 2.3.7A presents the estimated efficacy profile of each genotype group determined by SNP $rs1245576$ on chro-

mosome 10 and $rs1498716$ on chromosome 12, respectively. The estimated CE4 contrasts and their simultaneous confidence intervals (adjusted for both within- and across-SNP multiplicity) are also presented on the right panel. The efficacy profile for $rs1245576$ suggests a dominant beneficial effect of the minor allele, and it is confirmed by the contrast $\log \kappa_{2:(0,1)}$, of which the multiplicity-adjusted simultaneous confidence interval just misses covering 0. The $\{Aa, aa\}$ group by this SNP consists about 65.4% of the total population, a reasonable size to target. For the other SNP $rs1498716$, the efficacy profile suggests a recessive effect of the minor allele, and it is confirmed by the confidence intervals for the contrast $\log \kappa_{2:(0,1)}$. The targeted population $\{aa\}$ is about 19.6% of the entire population.

In this real data analysis, the results from CE4-Binary provides multiple SNPs from different regions as potential candidates for researchers to make their decisions. The final decision of which SNP or SNPs to be chosen as the biomarker for targeting requires multi-dimensional considerations. Note that the CE4 method does not directly consider the subgroups defined by more than one SNPs. As a post-hoc analysis, we further examined the targeted and non-targeted populations by the two SNPs illustrated above. Figure 2.3.7B presents the population into three categories: suggested as non-targeted population by both SNPs (M_0), suggested as targeted population by both SNPs (M_2), and suggested as non-targeted population by one SNP but targeted by the other SNP (M_1). It can be seen that more than half of the population (57%) belongs to M_1, which may appear to be inconclusive. Since these three groups co-defined by two SNPs may look similar to the situation where a single SNP separates the population into three genotype groups, one may consider applying CE4-Binary on this new 3-group scenario to further investigate the efficacy from each subgroup and their combinations. If that direction is pursued, one needs to understand that there is likely to have heterogeneous treatment efficacy within these newly defined subgroups, and the naive way of assuming the same efficacy for each subgroup may not be optimal (especially for M_1). As an alternative, one may consider to apply the general SME principle proposed by Ding et al. (2016) to cautiously get the efficacy estimates for each newly defined subgroup.

Finally, we examined the characteristics of targeted and non-targeted population based on each of the two SNPs (Table 2.3.6). The estimated relative risk is much lower in the

targeted group than that in the non-targeted group for both biomarkers (0.87 vs 1.85 for subgroups defined by $rs1245576$, and 0.58 vs 1.30 for subgroups defined by $rs1498716$). This is one of the advantages of the proposed CE4 method, where the corresponding estimated treatment efficacy in *both* the targeted and non-targeted group can be obtained together with a p-value for testing their difference. Other covariates including age, sex, smoking status and baseline severity score do not vary between the targeted group and non-targeted group, regardless of which SNP is used for targeting. It suggests that the differential effects are plausibly due to genetics.

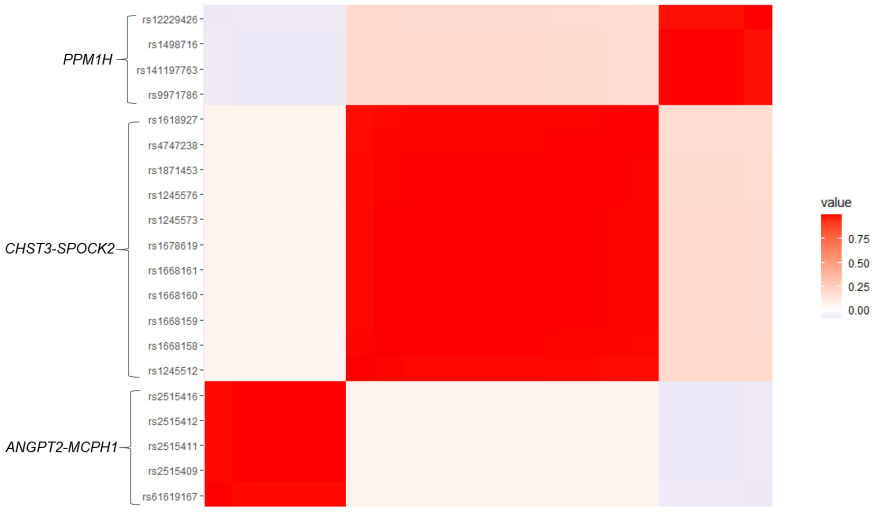

Figure 2.3.6: Heatmap of correlations between identified 20 SNPs.

2.3.5.3 Differential treatment efficacy persists in moderate to severe participants

As reported by Chew et al. (2013), the AREDS supplements was claimed to be most beneficial for moderate to severe patients, which corresponds to those in AMD categories 2, 3, and 4. We re-examined our findings by excluding the participants in category 1. Table 2.3.7 illustrated the characteristics for the patients among the two treatment arms. Note that since participants in category 1 were only allowed to be in the placebo arm, the number of participants taking the AREDS supplements remains the same. These participants were

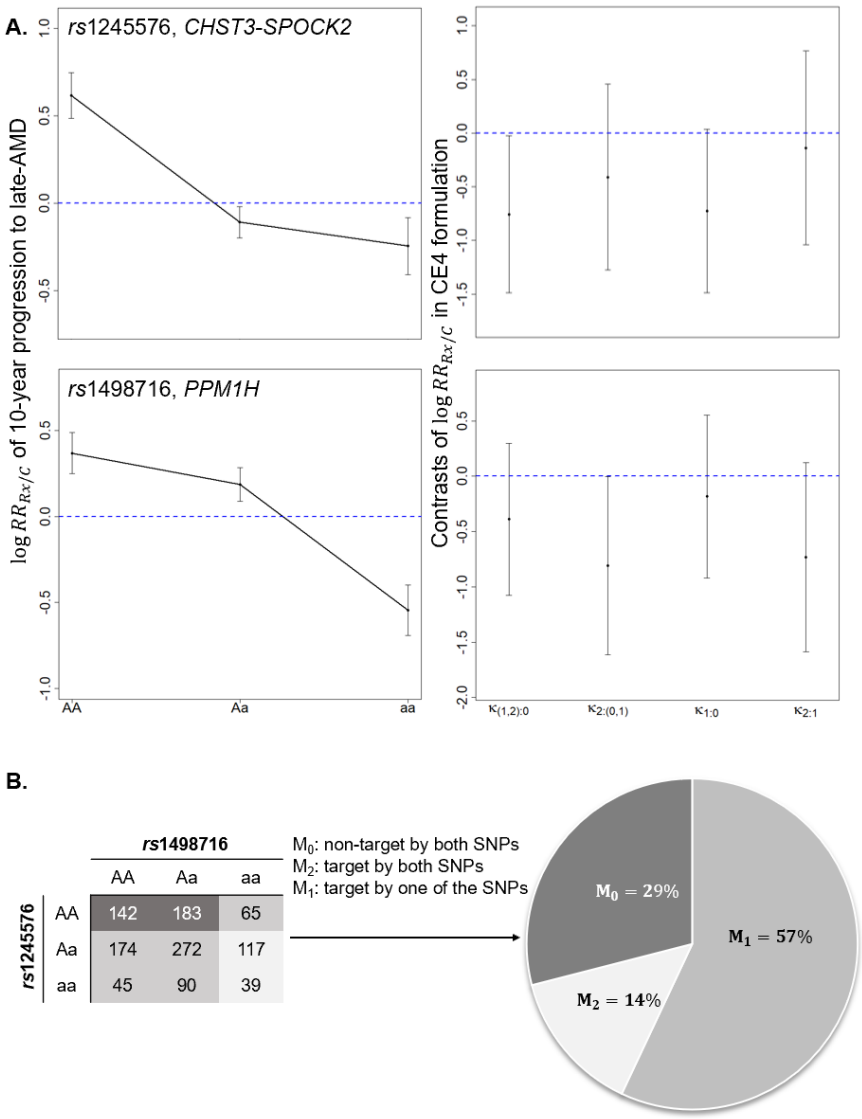

Figure 2.3.7: Two selected SNPs from AREDS analyses. **A:** treatment profile using log(RR) (left); CE4 estimates and their corresponding simultaneous confidence intervals (right). **B:** Sub-populations co-defined by two SNPs.

Table 2.3.6: Characteristics of targeted and non-targeted subgroups determined by two separate SNPs

	*rs*1245576: CHR10, *CHST3-SPOCK2*			*rs*1498716: CHR12, *PPM1H*		
	Targeted	Non-targeted	p-value	Targeted	Non-targeted	p-value
number of subjects (n,%)	737 (65.4)	390 (34.6)		221 (19.6)	906 (80.4)	
Treatment efficacy RR (SE)	0.87 (1.18)	1.85 (1.08)	$1.09 \times 10^{-6\star}$	0.58 (1.16)	1.30 (1.08)	$2.29 \times 10^{-6\star}$
Age			0.087			0.423
Mean (SD)	69.0 (4.8)	68.5 (4.9)		68.9 (4.9)	68.6 (4.9)	
Median (range)	68.1 (55.3-81.0)	68.9 (55.8-80.0)		68.8 (55.3-79.6)	68.9 (55.5-81.0)	
Sex (n, %)			0.194			0.845
Female	392 (53.2)	224 (57.4)		119 (53.8)	497 (54.9)	
Male	345 (46.8)	166 (42.6)		102 (46.2)	409 (45.1)	
Smoking (n, %)			0.745			0.550
Never Smoked	352 (47.8)	191 (49.0)		102 (46.2)	441 (48.7)	
Former/Current Smoker	385 (52.2)	199 (51.0)		119 (53.8)	465 (51.3)	
Treatment (n, %)			0.426			0.029
Placebo	436 (59.2)	241 (61.8)		118 (53.4)	559 (61.7)	
Antioxidant + Zinc	301 (40.8)	149 (38.2)		103 (46.6)	347 (38.3)	
Eye-level variables						
Number of eyes	n = 1399	n=705		n = 398	n=1646	
Baseline AMD severity score			0.530			0.130
Mean (SD)	3.2 (2.3)	3.1 (2.3)		3.3 (2.3)	3.1 (2.3)	
Median (range)	2.0 (1.0-8.0)	2.0 (1.0-8.0)		2.0 (1.0-8.0)	2.0 (1.0-8.0)	

$\star$: p-value is from the corresponding CE4 contrast when simultaneous type I error is controlled, without adjusting for cross-SNP multiplicity

randomized with equal probability to each treatment arm, and thus the baseline characteristics are all balanced between treatment and placebo. We specifically examined the two SNPs in the new population and obtained their efficacy profiles. Overall, they look almost identical as in the previous analyses. All the baseline characteristics do not differ between targeted and non-targeted population (Table 2.3.8), regardless of which SNP is used as the candidate marker for targeting. We conclude that the differential treatment efficacy that we discover persists in moderate to severe participants.

2.3.6 Summary and Discussion

Modern drug development in RCTs involves the evaluation of efficacy of a new treatment Rx relative to a control treatment C, based on a clinical meaningful outcome. This makes testing SNPs for a potential tailoring strategy fundamentally different from the traditional association detection for SNPs with a quantitative trait. Therefore, traditional genetic models or tests for association detection cannot be simply applied to such a SNP testing problem in drug development.

In this work, we develop a novel SNP testing method to confidently identify and infer subgroups with differential treatment efficacy from RCTs with binary outcomes. Our proposed CE4-Binary method, derived from the fundamental multiple testing principle, assesses all clinically plausible effects of a SNP through four contrasts. These contrasts are directly formulated based upon a logic-respecting efficacy measure, the RR. Using a log-linear model for the binary outcome, we demonstrate that the RR has a unique covariate-invariant property, which makes the comparison of treatment response between Rx and C straightforward in subgroups and their combinations.

Our multiplicity adjustment approach rigorously combines two error rate controls, strong FWER control within each SNP, and *per family* error rate control across the SNPs. Such an error control is appropriate for the new drug development purpose, which takes the dependence into account, both within each SNP and across the SNPs, and it allows flexibility in the exploration of multiple candidate SNPs, while being confident in the patient subgroup to target from any selected SNP.

Table 2.3.7: Characteristics of the AREDS participants with AMD category 2, 3, or 4

Number of subjects	Placebo (n=400)	Antioxidants and Zinc (n=450)	p-value*
Age			0.335
Mean (SD)	69.1 (5.1)	68.8 (5.0)	
Median (Range)	68.7 (55.3-81.0)	68.7 (55.5-79.8)	
Sex (n, %)			0.446
Female	218 (53.2)	258 (57.3)	
Male	182 (45.5)	192 (42.7)	
Smoking (n, %)			0.881
Never Smoked	195 (48.8)	216 (43.0)	
Former/Current Smoker	205 (51.2)	234 (52.0)	
AREDS AMD categories (n, %)			
1	0 (0)	0 (0)	
2	119 (29.8)	122 (27.1)	
3	190 (47.5)	225 (50.0)	
4	91 (22.8)	103 (22.9)	
Eye-level variables			
Number of eyes	n=702	n=789	
Baseline AREDS AMD severity score			0.517
Mean (SD)	4.0 (2.3)	3.9 (2.2)	
Median (Range)	4.0 (1.0-8.0)	4.0 (1.0-8.0)	

*p-value is based on two-sample t test or Chi-square test for continuous or categorical variables

Table 2.3.8: Characteristics of the targeted and non-targeted populations for participants with AMD category 2, 3, or 4

	*rs*1245576: CHR10, *CHST3-SPOCK2*			*rs*1498716: CHR12, *PPM1H*		
	Targeted	Non-targeted	p-value	Targeted	Non-targeted	p-value
# of subjects (n,%)	562 (66.1)	288 (33.9)		180 (21.2)	670 (78.8)	
Treatment efficacy RR (SE)	0.80 (1.08)	1.67 (1.14)	$1.89 \times 10^{-6\star}$	0.53 (1.16)	1.20 (1.08)	$1.19 \times 10^{-6\star}$
Age			0.277			0.379
Mean (SD)	68.8 (5.1)	69.2 (4.9)		69.2 (5.0)	68.8 (5.1)	
Median (range)	68.5 (55.3-81.0)	69.2 (55.8-80.0)		69.0 (55.3-79.6)	68.6 (55.5-81.0)	
Sex (n, %)			0.178			0.774
Female	305 (54.3)	171 (59.4)		103 (57.2)	373 (55.7)	
Male	257 (45.7)	117 (40.6)		77 (42.8)	297 (44.3)	
Smoking (n, %)			0.799			0.928
Never Smoked	274 (48.8)	137 (47.6)		86 (47.8)	325 (48.5)	
Former/Current Smoker	288 (51.2)	151 (52.4)		94 (52.2)	345 (51.5)	
Treatment (n, %)			0.666			0.226
Placebo	261 (46.4)	139 (48.3)		77 (42.8)	323 (48.2)	
Antioxidant + Zinc	301 (53.6)	149 (51.7)		103 (57.2)	347 (51.8)	
Eye-level variables						
Number of eyes	n = 990	n=501		n = 316	n=1175	
Baseline AMD severity score			0.835			0.693
Mean (SD)	3.9 (2.3)	4.0 (2.2)		3.9 (2.3)	4.0 (2.3)	
Median (range)	4.0 (1.0-8.0)	4.0 (1.0-8.0)		5.0 (1.0-8.0)	4.0 (1.0-8.0)	

$\star$: p-value is from the corresponding CE4 contrast when simultaneous type I error is controlled, without adjusting for cross-SNP multiplicity

We successfully applied CE4-Binary on AREDS data to identify subgroups that exhibit enhanced efficacy with the treatment of antioxidants plus zinc in reducing AMD progression rate. Multiple gene regions have been discovered to suggest subgroups with significantly enhanced efficacy, which include the *ANGPT2-MCPH1* region on chromosome 8, *CHST3-SPOCK2* on chromosome 10, and *PPM1H* on chromosome 12. Using two top SNPs as an example, we further examined the treatment efficacy and patient characteristics in the targeted and non-targeted subgroups. Our findings provide new perspectives on the differential treatment efficacy, suggested by genetic polymorphisms for reducing AMD progression rate.

Although we only focus on SNP testing in this article, the key elements of the method are applicable to broader scenarios with other types of markers for testing. For example, the marker separates the patient population into more groups (> 3) such as the immunohistochemistry test, or the subgroups co-defined by multiple markers (i.e., SNP and immunohistochemistry test). In these scenarios, additional contrasts need to be constructed to obtain the complete ordering of the treatment efficacy in individual subgroups and some of their combinations, which will then be used to identify the subgroups.

3.0 Individual Treatment Effects Estimation through Machine Learning in Survival Data

3.1 Introduction

One important aspect of precision medicine is allowing doctors to select treatments that are most likely to help patients based on their own clinical or other characteristics. Different from traditional clinical studies where the focus is on estimating the average treatment effect (ATE) in a representative population (usually through a well-designed clinical trial), assisting patients to shape their individualized-treatment plan requires an understanding of the heterogeneity of treatment effects (HTE) from a more patient-centric view. With the increase amount of large bioinformatic datasets and the use of electronic health record data, a full picture of individuals' characteristics is forming, which also brings challenges to statistical analyses due to the complexity of the data structure. Thus using flexible modeling techniques such as machine learning methods or deep learning methods within the counterfactual framework shows great potential and receives much attention. Foster et al. (2011a) proposed the virtual twins approach to study the treatment effect heterogeneity among individuals and then identify subgroups of individuals who can benefit from a treatment. The basic idea is to calculate the treatment effects by taking the difference between predicted response values obtained from an individual's observed and "twin" data point by altering the treatment assignment. The outcomes are then used in classification or regression trees on covariates to identify potential subgroups with differential treatment effects compared to the average value. Wager and Athey (2018) modified the random forest by changing the splitting rule to maximize the treatment difference within a node. They also use hold out data from different treatment groups to calculate the treatment effect for each terminal node and then average over the forest. Hill (2011) described an adaptive approach of the Bayesian Additive Regression Trees (BART) to accommodate the causal inference framework, which is similar to the virtual twin but instead using the BART as base learner. Lu et al. (2017) conducted simulations to compare the performance of different methods including virtual

twins, causal forest, synthetic forest and BART in estimating individual treatment effect (ITE). It is worth mentioning that in many research paper, the ITE refers to the conditional average treatment effect (CATE), which is not the true individual treatment effect, but rather the treatment effect for the observed individual that represents the cohort with the same set of characteristics. It is impossible to collect all information of an individual that could distinguish him or her from all other people, and thus the exchangeable use of ITE and CATE is acceptable.

Kunzel et al. (2018) formally defined the meta-algorithms to estimate the CATE function. A meta-algorithm consists two level of models. The base learner is used to build prediction model for the response value of individuals and it can be any form of machine learning methods, deep learning methods or the regression methods, which can easily be adapted to various types of outcome variables. The meta-level algorithm can be seen as a function of the base learners for computing the CATE of each individual. Three meta-algorithms were discussed in Kunzel et al. (2018) including: S-learner, T-learner and X-learner. The S-learner is constructed on a single prediction model where the treatment assignment is considered as one covariate similar to other characteristic variables. By altering the treatment assignment of an individual while keeping the rest covariates as observed values, the CATE is then computed as the difference between the two predicted values. The T-learner is based on two prediction models where each one is built on either the treatment or the control cohorts. In this setting, the treatment assignment is used to separate the population into two groups where two models are fit separately. The CATE is then estimated from taking the difference between the predicted values obtained from the two models. The X-learner is a modification of T-learner, which is more efficient when the treatment assignment is heavily unbalanced (i.e., one treatment arm contains much more patients than the other). After constructing two prediction models, the pseudo-CATE is imputed by taking the difference between observed outcome and the potential outcome obtained from the corresponding model. The imputed pseudo-CATE is then regressed on covariates to build two prediction models for each treatment arm (any type of machine learning/ deep learning/ regression model can be used) and thus for each individual two predicted "CATE" can be obtained. The two estimates are then combined using a weighted sum to get the final estimate of CATE. Details of the algorithms

are discussed in Section 1.5.

Time-to-event outcomes primarily arise from medical and biological studies and also widely exist in epidemiological, sociological, economic, and financial research. They are the most commonly used outcomes in cancer studies and thus for oncology drug development, understanding the estimation of CATE for survival outcomes is essential. However, the unique "missingness" happened in such outcomes which is known as censoring makes the survival analysis more complicated than other types of outcomes. While most studies we mentioned focus on the continuous or binary outcomes, there are some research about the estimation of CATE using time-to-event outcomes. Zhu and Gallego (2020) proposed a framework to first estimate ITE and then using a scoring system to identify variables contributing to the HTE. They converted the original survival data into a sequence of binary outcomes over time and then applied the Super-learner to estimate the conditional hazard rate at each time point. The estimation step is more on the binary type of data rather than the survival data. Cui et al. (2020) recently developed a causal survival tree method which is an extension of Wager and Athey (2018) to accommodate right censoring in survival outcomes. Tabib and Larocque (2020) proposed to use a special splitting rule in the random forest where the distance between the ITE of the left and right node population is maximized considering the node sizes. Unfortunately, none of the three methods has a publicly available software or package.

In this project, we specifically examined the use of the meta-algorithm with machine learning base learners that are designed for survival outcomes to estimate CATE for each individual. Further, based on the CATE estimation, we will identify potential subgroups for recommending appropriate treatment. We will identify the prescriptive variables that predicts the HTE to help understand the mechanism of the HTE. The chapter is organized as follows. Section 3.2 introduces the framework of the problem, treatment efficacy measures to use, and the machine learning/deep learning methods we considered, followed by an algorithm to identify important variables related to the treatment recommendation. In Section 3.3, intensive simulations are conducted to compare the performance of different models in estimating CATE and recommending the "right" treatment for individuals under various conditions. The models are then applied to the AREDS data in Section 3.4 and important variables contributing to the treatment recommendation are identified. Finally,

we discuss and conclude in Section 3.5.

3.2 Methods

3.2.1 Framework and notations

Consider a study with two treatment arms and denote $Z_i \in \{0, 1\}$ a binary treatment assignment variable ($Z_i = 1$ for treatment and $Z_i = 0$ for control). $\boldsymbol{X_i}$ is a $p-$dimensional covariate vector. Let $(T_i(0), T_i(1))$ denote the counterfactual survival times of individual i under Neyman-Rubin potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990). If the observed survival time and censoring time are (T_i, C_i), then the observed outcome would be $Y_i = \min(T_i, C_i)$ with the event indicator $d_i = I(T_i < C_i)$. We assume the censoring time is independent of the survival time, which is also known as noninformative censoring: $C_i(z) \perp\!\!\!\perp T_i(z)|(\boldsymbol{X_i}, Z_i)$. To aid the estimation of CATE, the following assumptions need to be made.

Assumption 1 (Consistency)

$$T_i = Z_i T_i(1) + (1 - Z_i) T_i(0),$$

Assumption 2 (Unconfoundedness)

$$Z_i \perp\!\!\!\perp (T_i(0), T_i(1))|\boldsymbol{X_i},$$

Assumption 3 (Population Overlap)

$$P(Z_i = 1|\boldsymbol{X_i} = \boldsymbol{x_i}) \in (0, 1).$$

The consistency assumption guarantees the counterfactual model applied to the observed outcomes as: $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ and $d_i = Z_i d_i(1) + (1 - Z_i) d_i(0)$. The unconfoundedness assumption requires the treatment assignment to be independent of the potential outcomes given covariates sets, which rules out the existence of unobserved factors that affect treatment choice and are also correlated with the outcomes. The population overlap

assumption describes that for each value of covariate set, there is a positive probability of being assigned to both treatment and control arms, or equivalently, there is sufficient overlap in the characteristics of treated and untreated patients for adequate matches.

The commonly used treatment efficacy measures for time-to-event outcomes include difference or ratio of: (1) survival probability at a specific timepoint, (2) mean (restricted) survival time, and (3) quantile survival time. We specifically examined the survival probability at a specific timepoint in the project, which is defined as $S(t_i = t|Z_i, \boldsymbol{X_i}) = P(t_i > t|Z_i, \boldsymbol{X_i})$. The corresponding CATE function can then be determined as

$$\theta(t, \boldsymbol{x}) = E[S(t|1, \boldsymbol{x}) - S(t|0, \boldsymbol{x})] \tag{3.2.1}$$

3.2.2 Estimating CATE with meta-algorithms

The estimation of CATE largely rely on the performance of base learners when using the meta-algorithms. Machine learning methods stand out when complex high-dimentional data become available nowadays. For survival outcomes, we consider using three machine learning/ deep learning techniques as base learners: random survival forest (RSF, (Ishwaran and Kogalur, 2007)), Bayesian accelerated failure time model (BAFT, (Henderson et al., 2020)), and deep neural network (DNN) survival model (Sun et al., 2020).

We first review the S-learner using the survival outcomes. A single prediction model is fitted using observed outcomes and covariates, together with the treatment assignment, denoted as $\hat{S}(t, Z, X) =$. The predicted survival probability at a given timepoint under different treatment conditions is then obtained by altering the treatment assignment for individuals and the CATE estimate is then given by

$$\hat{\theta}(t, \boldsymbol{x}) = \hat{S}(t, 1, \boldsymbol{x}) - \hat{S}(t, 0, \boldsymbol{x}).$$

As noted in Foster et al. (2011a), manually including treatment-covariate interactions in the design matrix can improve the performance of the tree-based learners or algorithms. Including the interaction terms, ZX and $(1 - Z)X$, in the S-learner using RSF as the base learner adds the chance that the treatment assignment is selected in the variable sets to grow

a tree. Since the BART algorithm is in general more adaptive, there is no need to manually add the interaction matrix.

As for the T-learner, two prediction models are fitted using individuals under different treatment conditions. In this case, the treatment indicator is no longer a covariate in the prediction model. Denote $\hat{S}_1(t, X)$ as the prediction model using individuals with treatment, and $\hat{S}_0(t, X)$ as the prediction model using individuals with control. The predicted survival probability at a given time point under different treatment conditions is then obtained by fitting the two models using the same set of covariates from individuals and the CATE estimate is then given by

$$\hat{\theta}(t, \boldsymbol{x}) = \hat{S}_1(t, \boldsymbol{x}) - \hat{S}_0(t, \boldsymbol{x}).$$

The X-learner is based on T-learner but additional steps are involved to have a more efficient estimate when unbalanced design happens (i.e., much more individuals are randomized in one treatment condition than that in the other condition). Since the individual-level survival probability at certain time point is not observed, the imputed treatment effects for individuals in the treated group and control group are modified as:

$$\tilde{D}^1(\boldsymbol{x}^1) = \hat{S}_1(t, \boldsymbol{x}^1) - \hat{S}_0(t, \boldsymbol{x}^1),$$
$$\tilde{D}^0(\boldsymbol{x}^0) = \hat{S}_1(t, \boldsymbol{x}^0) - \hat{S}_0(t, \boldsymbol{x}^0).$$

Here the superscripts $\{0, 1\}$ denote the cohort under control or treatment condition. Any supervised learning or regression method(s) can be applied to estimate the treatment effects using either the imputed treatment effects in the treatment group to obtain $\hat{\theta}_1(\boldsymbol{x}) = E[\tilde{D}^1|X^1 = x]$ and similarly in the control group to obtain $\hat{\theta}_0(\boldsymbol{x}) = E[\tilde{D}^0|X^0 = x]$. Thus for each individual there are two potential estimates for CATE. The final CATE estimate is defined as a weighted sum of the two estimates in this stage:

$$\hat{\theta}(\boldsymbol{x}) = g(\boldsymbol{x})\hat{\theta}_0(\boldsymbol{x}) + (1 - g(\boldsymbol{x}))\hat{\theta}_1(\boldsymbol{x}),$$

where $g \in (0, 1)$ is a weight function. Following the recommendation of Kunzel et al. (2018), we use the estimate for the propensity score $P[Z = 1|\boldsymbol{X}]$ as $g(\boldsymbol{x})$.

As pointed out by Kunzel et al. (2018), the performance of each estimate can largely vary by the settings. For example, since the S-learner treats the treatment assignment like other predictors, without considering modifying the model by including the interaction terms, the original model using RSF as base learner can completely ignore the treatment indicator by not splitting on it. This only works well when the CATE is around 0 (i.e., no treatment effects), which is not of our interest. Adding interactions can solve the issue, but the estimates can be biased when the treatment assignment is severely unbalanced.

3.2.3 Identification of important variables contributing to treatment recommendation rule

CATE describe the heterogeneity of treatment effects. The next step is to use CATE to select the "correct" treatment for each individual, i.e., to build the treatment recommendation rule. A straightforward approach is to use 0 as the threshold (when the treatment effects are based on the difference of survival probability). Patients with estimated CATE greater than zero would benefit more from taking the treatment while for those with estimated CATE less than zero, the control is the to-go drug. This way the treatment recommendation for patients becomes a binary classification problem, and the important variables are identified as those that contribute most to minimize the classification error. We used Boruta algorithm (Kursa and Rudnicki, 2010) for identification of these important variables.

The algorithm started by generating "shadow variables" by shuffling the values of each original feature to remove any potential correlation with the target binary outcome, in our case, the treatment recommendation. These "shadow variables" are then added to the original dataset so that the total number of variables is doubled. A random forest classifier is used to run on the extended dataset and z-scores are computed for all variables including real and shadow ones. The maximum score of all "shadow variables" serves as a threshold to assign a "Hit" to the original variables (i.e., if the z-score of original feature is greater than the threshold, a "Hit" is marked). For variables with undetermined importance, a two-sided test of equality with the max z-score in the shadow variables is conducted and those with significantly lower z-score are permanently removed from the dataset. After removing

unimportant variables and all shadow variables, repeat the procedure until either the importance is assigned for all variables or the pre-specified number of iterations is reached. This algorithm identifies variables that significantly contribute to the treatment recommendation strategy (i.e., whether the individual can benefit from one treatment over the other treatment). We can understand the heterogeneity in the treatment efficacy from these important variables and provide guidance for future trial designs.

3.3 Simulation study

3.3.1 Simulation design

We conducted intensive simulations to compare the finite sample performance of the proposed methods for estimating CATE. We simulated 10 independent covariates $X_1, \ldots, X_p$ from $N(0, 0.35^2)$ and dichotomized X_8, X_9, X_{10} through $I(X > 0)$. The total sample size was $n = 1,000$. The treatment indicator was generated from $Z|X \sim \text{Bernoulli}(\exp(\boldsymbol{X}))$, where $Z = 1$ for treatment group and $Z = 0$ for control group. We considered three randomization designs: (1) balanced design where $\exp(\boldsymbol{X}) = 0.5$; (2) unbalanced design where $\exp(\boldsymbol{X}) = 0.05$; and (3) slight violation of unconfoundedness assumption where $\text{logit}(\exp(\boldsymbol{X})) = \beta_0 + 1.3 \times X_1 - 0.8 \times X_5$. β_0 was selected to have an overall $1:1$ treatment-control ratio. We denote the third design as dependent design in the following context.

The survival times T were simulated from the following Weibull survival curve:

$$S(t|\boldsymbol{X} = \boldsymbol{x}, Z = z) = \exp\left[\exp\{h_z(\boldsymbol{x})\}(\frac{t}{\lambda_z})^{\eta}\right],$$

where $h_z(\boldsymbol{x})$ denotes the prognostic function under Cox model. By inverse-transforming the survival function, the survival times were then simulated as:

$$T(\boldsymbol{X} = \boldsymbol{x}, Z = z) = \lambda_z \left[\frac{-\log(U)}{\exp\{h_z(\boldsymbol{x})\}}\right]^{\frac{1}{\eta}},$$

where $U \sim \text{Unif}[0, 1]$. We set the shape parameter $\eta = 2$ and the scale parameter $\lambda_0 = 18$ and $\lambda_1 = 20$ so that individuals in the treatment group have a larger baseline ($h_z(x) = 0$)

survival probability, i.e., longer survival time. The censoring times were simulated from an exponential distribution to have an overall censoring rate of 30%. Three types of heterogeneous treatment effects were introduced by the following structures for $h_z(\boldsymbol{x})$:

$$\begin{aligned}
\text{Scenario 1: } h_0(\boldsymbol{x}) &= 0.2X_1 + 0.7X_2 + 0.4X_9, \\
h_1(\boldsymbol{x}) &= -0.5X_1 - 2X_2 - 0.25X_9. \\
\text{Scenario 2: } h_0(\boldsymbol{x}) &= -0.5X_1 + 0.7X_2 + 0.2X_9 + 0.9X_2X_9, \\
h_1(\boldsymbol{x}) &= -0.05e^{X_1} - 0.2X_2^2 + 0.35X_9. \\
\text{Scenario 3: } h_0(\boldsymbol{x}) &= -0.5X_1 + 0.7X_2 + 0.2X_9 + 0.9X_2X_9 + 0.6X_3 - 0.5X_4^2 + 0.6X_8, \\
h_1(\boldsymbol{x}) &= -0.05e^{X_1} - 0.2X_2^2 + 0.2X_9 - 0.1e^{X_5} + 0.7\sin(X_6) + 0.5X_{10}.
\end{aligned} \tag{3.3.1}$$

From this setting, the ITEs are determined by X_1, X_2, and X_9 in a linear form in Scenario 1, and in a complex form in Scenario 2. In Scenario 3, treatment effects of the two groups are functions of two different sets of covariates, and two treatment groups share a small subset of the covariates ((X_1, X_2, and X_9). We explored the following combinations of the meta-algorithms and base learners: T-learner with RSF (R-T), X-learner with RSF (R-X), T-learner with BAFT (B-T), X-learner with BAFT (B-X), T-learner with DNNSurv (D-T) and X-learner with DNNSurv (D-X). A true Weibull model was also included as the optimal CATE estimate. We trained each model on the simulated training set of $n = 1,000$ patients, and we evaluated its performance on an independent test set of $N = 10^5$ patients for which we can calculate the true CATE from their covariate sets. We repeated the Monte-Carlo simulation $B = 100$ times. The performance measures we assessed are bias and root mean squared error (RMSE) within quantile-bins which are defined as:

$$\text{Bias} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{n_q}\sum_{j=1}^{n_q}(\hat{\theta}_j - \theta_j),$$

$$\text{RMSE} = \frac{1}{Q}\sum_{i=1}^{Q}\sqrt{\frac{1}{n_q}\sum_{j=1}^{n_q}(\hat{\theta}_j - \theta_j)^2}.$$

Here Q denote the number of bins based on true CATE quantiles, where true CATE for individual j was computed by:

$$\theta_j = S_j(t = \tilde{t}|\boldsymbol{X} = \boldsymbol{x_j}, 1) - S_j(t = \tilde{t}|\boldsymbol{X} = \boldsymbol{x_j}, 0) = e^{[\exp\{h_1(\boldsymbol{x_j})(\frac{\tilde{t}}{\lambda_1})^\eta\}]} - e^{[\exp\{h_0(\boldsymbol{x_j})\}(\frac{\tilde{t}}{\lambda_0})^\eta]}.$$

Table 3.3.1: 2×2 table of recommendation rule from true CATE and estimated CATE

	Rule by θ	
Rule by $\hat{\theta}$	TP	FP
	FN	TN

The overall median time for all individuals in the test dataset was used as $\tilde{t}$, and other parameters from (3.3.1) were plugged in the above equation, together with the covariate set of individual j. We set $Q = 50$ in the simulation studies. For X-learner, the machine learning method used for estimating imputed treatment effects was gradient boosting machine (GBM) (Greenwell et al., 2020), and the estimate of propensity score was obtained from random forest (RF) classification (Liaw and Wiener, 2002).

The CATE estimate from each method was used to set up treatment recommendation rule. Patients with $\hat{\theta} > 0$ were labeled as recommended for treatment (RT) and patients with $\hat{\theta} < 0$ were labeled as recommended for control (RC). Assuming that the label based on true CATE is the gold standard, the recommendation rule generated by each CATE estimate is a classification problem and the prediction accuracy metrics for 2×2 table can be used (Table 3.3.1). The following measurements were considered: accuracy ($ACC = \frac{TP+TN}{N}$), positive predictive value ($PPV = \frac{TP}{TP+FP}$), negative predictive value ($NPV = \frac{TN}{TN+FN}$), sensitivity ($= \frac{TP}{TP+FN}$), specificity ($= \frac{TN}{TN+FP}$), and F-score ($= \frac{2}{PPV^{-1}+\text{Sensitivity}^{-1}}$).

3.3.2 Simulation results

3.3.2.1 Balanced Design

Figure 3.3.1 presents the binned bias and RMSE for CATE estimates from each method under balanced design. The CATE estimate from correctly specified Weibull model shows the best performance one can get. In general, the BAFT base learners (i.e., B-T and B-X) have relatively large biases and RMSEs compared to other methods, except for the simple

linear case where they have the smallest biases. The RSF- and DNNSurv-based models have comparable biases under all scenarios, but DNNSurv-models provide smaller RMSE. Overall, the X-learner has a better performance in terms of RMSE than T-learner. The prediction accuracy metrics are summarized in Table 3.3.2, with the best performer other than Weibull model being highlighted in bold. For scenario 1 when only linear terms are involved, X-learner with DNNSurv had the best measurements across all six metrics. For scenario 2 and 3, X-learner with DNNSurv provides the best or the second best performance for all metrics. While T-learner with DNNSurv and X-learner with RSF outperform D-T at some metrics under more complicated scenarios (e.g., D-T has a larger specificity under both scenario 2 and 3, and PPV under scenario 3), the D-X still provides comparable prediction accuracy.

3.3.2.2 Unbalanced Design

Under unbalanced design, the BAFT-based learners have smallest biases under relatively simple scenarios (scenarios 1 and 2) as compared to other methods, but still show inflated RMSEs compared to RSF-based methods. The biases from RSF- and DNNSurv-based models are similar across all scenarios, but the RMSEs from DNNSurv-based methods show much larger biases (Table 3.3.2). In the simple linear scenario, D-X has the best ACC, PPV and specificity whereas R-T has the best NPV, sensitivity and F-score. In Scenario 2 when non-linear terms are involved R-T outperforms other methods for all metrics except for specificity where it has the second best performance after B-T. In the most complex scenario (Scenario 3), no single model outshines others. R-T, R-X and B-T are the best performer under different metrics.

With two hidden layers and 30 nodes per layer, DNNSurv-based models contain more tuning parameters than other machine learning methods, and thus require sizeable samples to obtain a stabilized and minimized loss value. In the unbalanced design with $n = 1,000$ and $e(X) = 0.05$, the treatment arm only contains about 50 people and around 30% are censored. Such a small sample size can cause variable results across 100 runs. To demonstrate the impact of sample size, we combined 2, 4, and 10 training datasets to enlarge the training

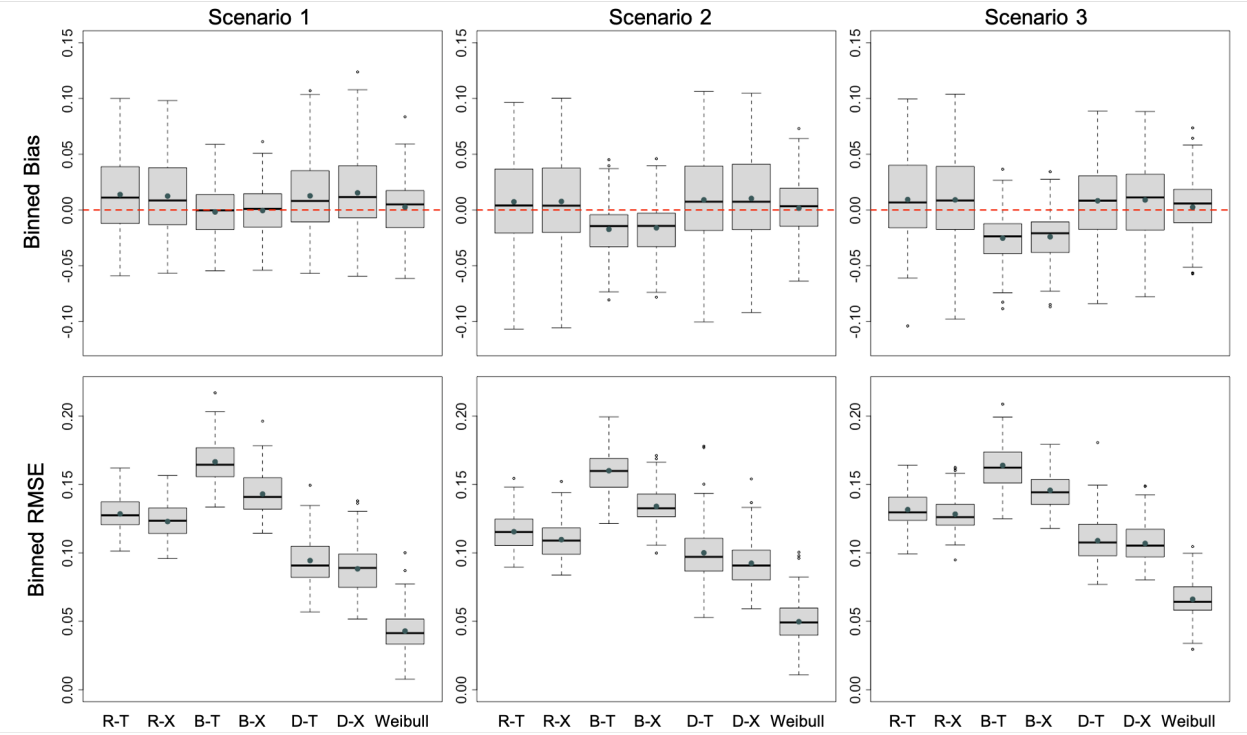

Figure 3.3.1: Box plots to compare the performance of CATE estimates: balanced design.

Table 3.3.2: Prediction Accuracy using CATE estimates: balanced design

	R-T	R-X	B-T	B-X	D-T	D-X	Weibull
Scenario 1							
ACC mean (SD)	90.11 (1.81)	90.58 (1.85)	85.99 (2.03)	88.55 (1.91)	92.39 (1.71)	**93.44 (1.67)**	96.37 (1.62)
PPV mean (SD)	91.78 (3.21)	92.33 (3.30)	90.67 (2.32)	92.18 (2.50)	94.96 (3.02)	**95.52 (2.94)**	97.36 (2.33)
NPV mean (SD)	87.17 (5.99)	87.56 (6.19)	76.13 (4.54)	80.93 (4.81)	87.62 (5.77)	**89.63 (5.43)**	94.68 (4.90)
Sensitivity mean (SD)	94.56 (3.24)	94.62 (3.38)	89.26 (2.96)	91.51 (2.94)	94.30 (3.25)	**95.25 (3.04)**	97.55 (2.44)
Specificity mean (SD)	79.74 (9.21)	81.16 (9.29)	78.36 (6.22)	81.64 (6.66)	87.92 (8.00)	**89.20 (7.63)**	93.62 (5.89)
F-score mean (SD)	93.05 (1.22)	93.36 (1.27)	89.90 (1.52)	91.79 (1.39)	94.54 (1.22)	**95.31 (1.20)**	97.41 (1.15)
Scenario 2							
ACC mean (SD)	79.68 (4.01)	80.50 (4.15)	72.15 (3.78)	75.29 (3.91)	82.10 (4.50)	**82.76 (4.68)**	91.04 (3.53)
PPV mean (SD)	86.55 (4.87)	86.97 (5.11)	84.69 (3.50)	86.18 (4.01)	89.60 (5.28)	**89.63 (5.78)**	94.07 (4.47)
NPV mean (SD)	66.96 (9.64)	68.80 (10.19)	51.73 (5.27)	56.37 (6.08)	70.15 (11.31)	**72.14 (11.67)**	86.84 (10.55)
Sensitivity mean (SD)	85.31 (8.21)	86.15 (8.26)	74.42 (5.21)	78.00 (5.61)	85.38 (8.42)	**86.57 (8.68)**	93.69 (6.01)
Specificity mean (SD)	65.89 (16.12)	66.65 (16.80)	66.60 (9.65)	68.66 (11.45)	**74.07 (16.28)**	73.41 (18.48)	84.53 (12.63)
F-score mean (SD)	85.50 (3.47)	86.12 (3.51)	79.08 (3.20)	81.70 (3.19)	87.00 (3.74)	**87.57 (3.78)**	93.64 (2.65)
Scenario 3							
ACC mean (SD)	83.47 (2.16)	83.74 (2.20)	79.21 (2.83)	81.30 (2.57)	86.73 (2.58)	**86.77 (2.19)**	92.12 (1.89)
PPV mean (SD)	86.61 (3.66)	86.42 (3.86)	87.52 (2.46)	88.06 (2.57)	**91.22 (3.19)**	90.34 (3.70)	94.86 (2.53)
NPV mean (SD)	76.62 (7.85)	78.17 (7.93)	62.53 (5.16)	66.87 (5.52)	77.89 (8.26)	**79.66 (7.75)**	86.67 (6.79)
Sensitivity mean (SD)	91.33 (4.92)	**92.08 (4.77)**	82.76 (4.15)	85.53 (4.31)	90.35 (5.21)	91.58 (4.73)	94.20 (3.63)
Specificity mean (SD)	63.91 (12.66)	62.95 (13.52)	70.37 (7.06)	70.77 (7.78)	**77.73 (9.68)**	74.77 (11.60)	86.96 (7.14)
F-score mean (SD)	88.72 (1.59)	88.97 (1.50)	84.99 (2.28)	86.67 (2.06)	90.62 (2.05)	**90.78 (1.60)**	94.45 (1.41)

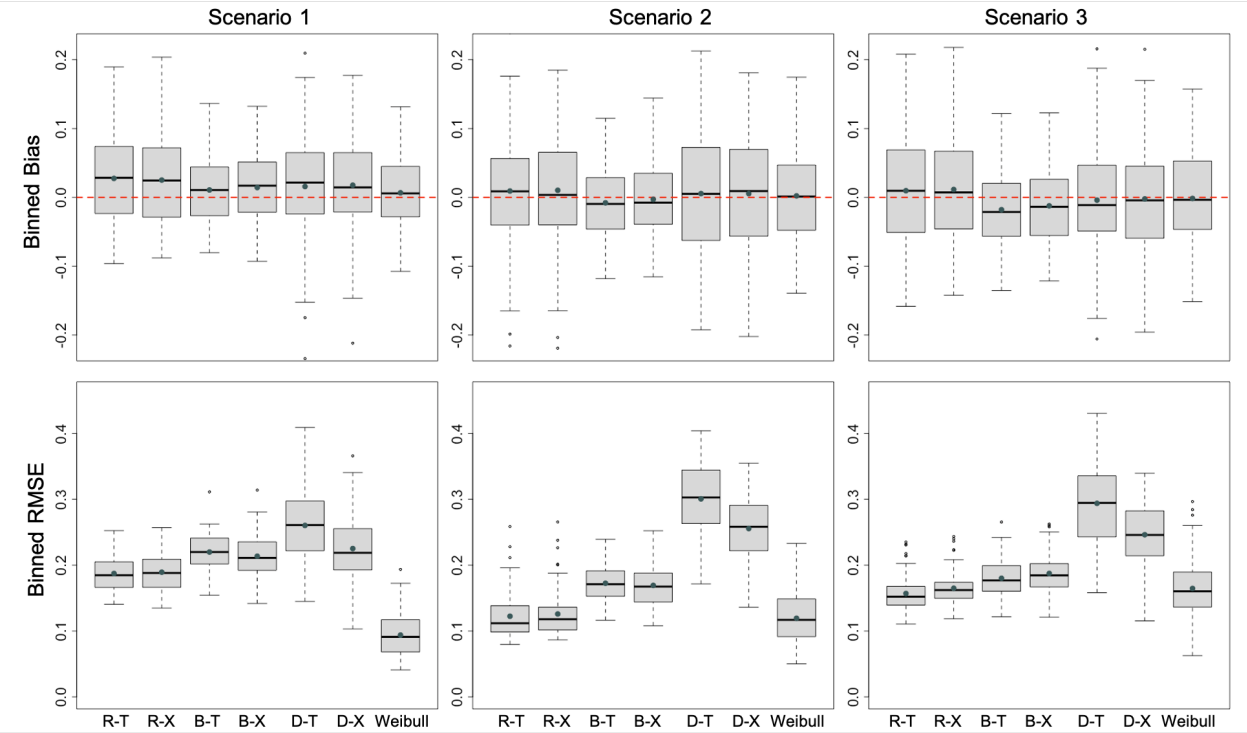

Figure 3.3.2: Box plots to compare the performance of CATE estimates: unbalanced design.

Table 3.3.3: Prediction Accuracy using CATE estimates: unbalanced design

	R-T	R-X	B-T	B-X	D-T	D-X	Weibull
				Scenario 1			
ACC mean (SD)	81.38 (5.73)	81.28 (6.74)	78.90 (3.79)	79.79 (4.80)	78.74 (6.20)	**81.56 (5.78)**	91.29 (3.59)
PPV mean (SD)	81.74 (6.87)	82.16 (8.19)	83.20 (4.33)	83.48 (5.31)	88.02 (4.75)	**89.02 (5.31)**	94.13 (4.73)
NPV mean (SD)	**88.55 (11.52)**	87.31 (11.31)	68.91 (9.07)	71.31 (10.76)	63.95 (10.32)	68.88 (10.54)	87.27 (9.38)
Sensitivity mean (SD)	**95.99 (5.50)**	95.58 (5.40)	88.07 (5.80)	89.43 (5.85)	80.96 (8.95)	84.53 (8.09)	93.82 (5.31)
Specificity mean (SD)	47.33 (24.57)	47.94 (29.10)	57.52 (14.47)	57.31 (17.88)	73.57 (12.32)	**74.62 (14.47)**	85.40 (12.85)
F-score mean (SD)	**87.94 (3.29)**	87.91 (3.72)	85.36 (2.72)	86.11 (3.19)	83.99 (5.39)	86.38 (4.75)	93.77 (2.59)
				Scenario 2			
ACC mean (SD)	**76.52 (7.74)**	75.61 (8.11)	71.42 (6.09)	70.18 (6.80)	64.10 (8.58)	65.78 (8.51)	80.01 (10.25)
PPV mean (SD)	**86.51 (7.18)**	86.14 (8.13)	83.99 (4.65)	82.42 (5.58)	77.89 (6.10)	79.39 (6.56)	88.75 (6.15)
NPV mean (SD)	**65.65 (14.14)**	63.54 (15.07)	52.01 (9.11)	50.20 (9.92)	41.43 (11.85)	44.19 (12.58)	71.22 (19.75)
Sensitivity mean (SD)	**81.55 (16.47)**	81.01 (17.09)	74.32 (10.03)	74.53 (11.24)	69.39 (12.12)	70.82 (13.19)	83.07 (16.18)
Specificity mean (SD)	64.20 (25.16)	62.39 (28.57)	**64.32 (13.76)**	59.51 (17.65)	51.13 (17.40)	53.41 (19.92)	72.52 (18.47)
F-score mean (SD)	**82.20 (9.82)**	81.48 (10.75)	78.38 (5.82)	77.64 (6.50)	72.83 (8.08)	74.05 (8.23)	84.71 (9.73)
				Scenario 3			
ACC mean (SD)	**78.71 (3.75)**	77.37 (3.89)	77.16 (4.71)	75.53 (4.61)	69.76 (6.83)	73.47 (6.03)	81.38 (5.76)
PPV mean (SD)	84.60 (6.22)	82.77 (6.09)	**85.99 (4.18)**	84.07 (4.20)	81.62 (5.35)	83.85 (5.14)	88.94 (4.76)
NPV mean (SD)	70.18 (11.67)	**70.50 (13.25)**	60.60 (8.92)	58.86 (9.55)	48.97 (11.19)	54.99 (11.35)	68.90 (12.97)
Sensitivity mean (SD)	87.40 (10.45)	**88.01 (10.99)**	81.67 (7.92)	81.67 (8.74)	74.79 (9.61)	78.49 (9.83)	84.93 (9.22)
Specificity mean (SD)	57.07 (23.29)	50.87 (24.46)	**65.92 (13.48)**	60.25 (14.82)	57.22 (15.43)	60.99 (16.52)	72.55 (14.72)
F-score mean (SD)	**85.20 (3.62)**	84.49 (3.88)	83.46 (4.05)	82.45 (4.14)	77.67 (5.90)	80.59 (5.32)	86.48 (4.80)

sample size into 2,000, 4,000, and 10,000, with treatment arm roughly contains 100, 200, and 500 individuals, and the performance was evaluated on the test dataset. The number of replications was deducted from 100 to 50, 25 and 10 respectively. Figures 3.3.3, 3.3.4, and 3.3.5 present the biases and RMSEs of CATE estimates on test dataset, with model trained by the combined training datasets. Firstly, as sample size increases, the variation of the estimates is getting smaller, reflected by narrow boxes. When 2 training datasets are combined, the RMSEs from D-T and D-X models are comparable to the BAFT-based model, but still greater than the RSF-based model. As sample size increases to 200 individuals in the treatment arm (combining 4 training datasets), the RMSEs from DNNSurv-based models are similar to those from RSF-based models in scenarios 2 and 3, and in scenario 1, the RMSEs are the smallest. Finally, when 10 datasets are combined, DNNSurv-based models become the best performers with the smallest RMSEs. Note that results on Figure 3.3.5 are from only 10 runs, so it is understandable that the biases are relatively large.

Under unbalanced design, the X-learner is expected to have the best performance since it adjusts the estimate based on the unbalanced allocation. However, we do not see a clear trend in the modified version for the time-to-event outcomes. More discussions are in Section 3.5.

3.3.2.3 Dependent Design

We further examined the performance under slight violation of unconfoundedness assumption where the treatment assignment depends on covariates X_1 and X_5. The overall allocation was 1:1. The CATE estimates had similar performance as compared to the balanced design regarding biases and RMSEs (Figure 3.3.6), indicating the robustness of these methods under minor violation of the unconfoundedness assumption. The D-T and D-X still provided the best estimates with small biases and RMSEs, followed by RSF-based models. Similar conclusions can be drawn from the prediction accuracy metrics. According to Table 3.3.4, D-T and D-X in general have comparable results and they tend to have the best predication accuracy. While R-X is the best performer for some metrics under different scenarios (NPV in scenario 1, and sensitivity in all three scenarios).

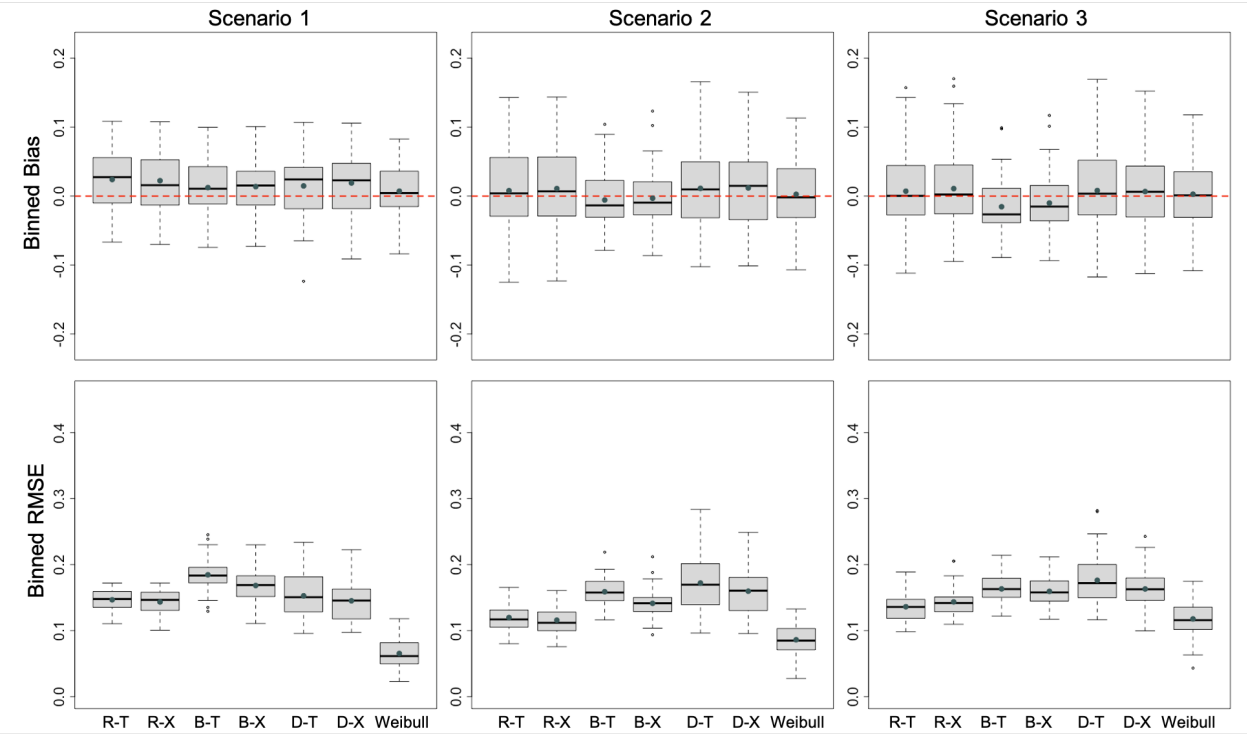

Figure 3.3.3: Box plots to compare the performance of CATE estimates: unbalanced design, combining 2 training datasets.

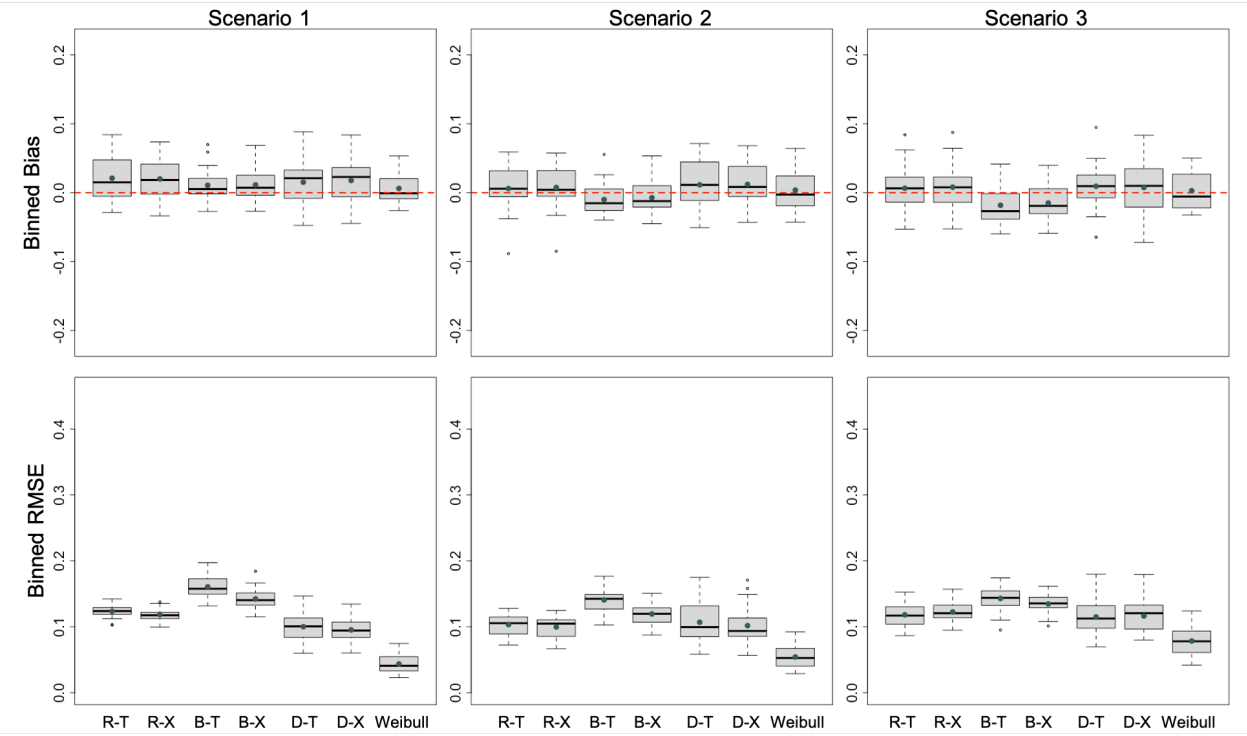

Figure 3.3.4: Box plots to compare the performance of CATE estimates: unbalanced design, combining 4 training datasets.

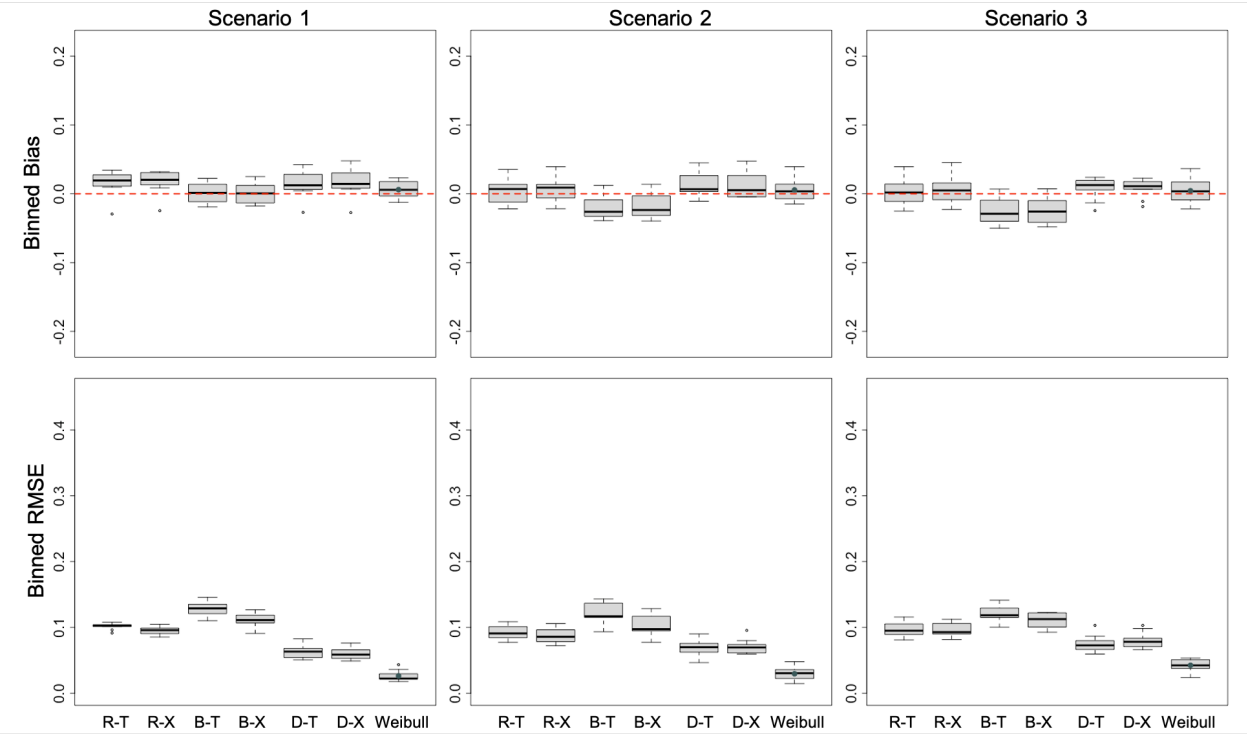

Figure 3.3.5: Box plots to compare the performance of CATE estimates: unbalanced design, combining 10 training datasets.

In summary, from simulation studies, DNNSurv-based models tend to provide more reliable estimates of CATE and the recommendation rule proposed based on the estimates are closer to the classification rule based on the true CATE. Results from T-learner and X-learner are similar most of the times. DNNsurv-based model may suffer from a limited sample size, where RSF and BAFT perform reasonably. All methods have shown some robustness against minor violation of the unconfoundedness assumption.

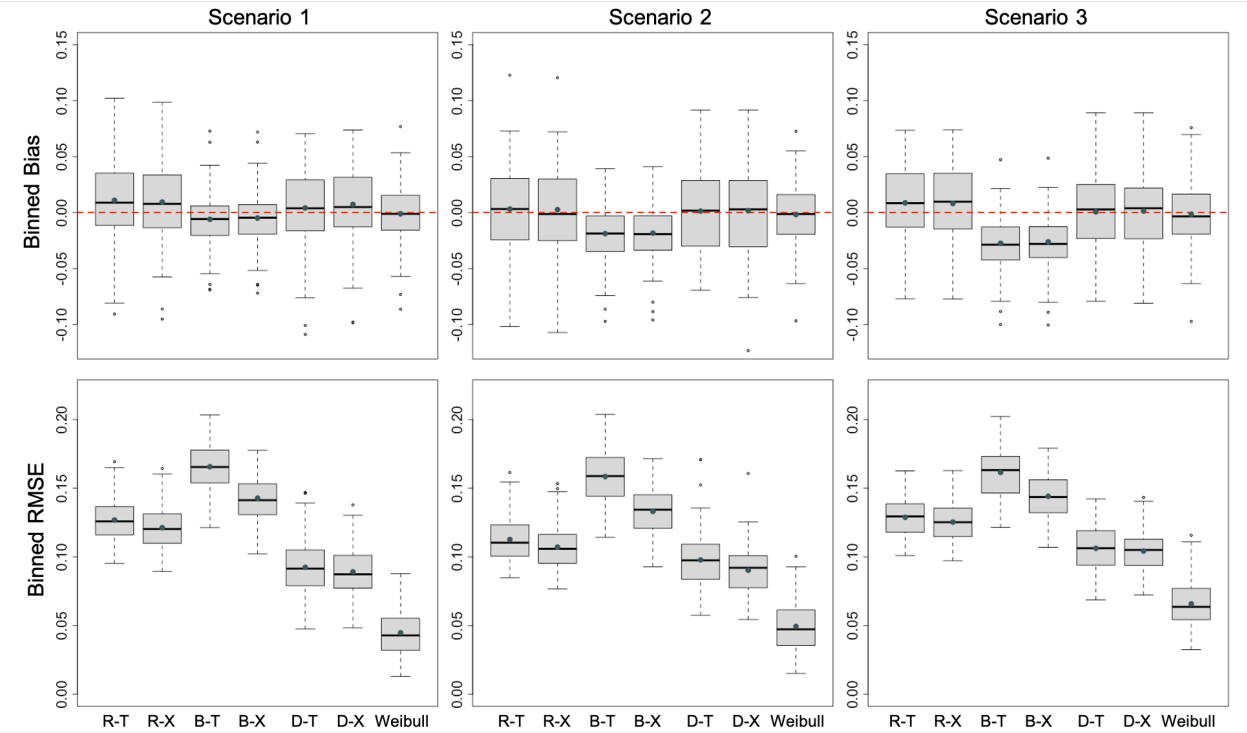

Figure 3.3.6: Box plots to compare the performance of CATE estimates: Z depends on X.

Table 3.3.4: Prediction Accuracy using CATE estimates: Z depends on X

	R-T	R-X	B-T	B-X	D-T	D-X	Weibull
Scenario 1							
ACC mean (SD)	89.86 (1.67)	90.41 (1.72)	85.64 (2.18)	88.21 (1.94)	92.23 (2.34)	**92.99 (2.25)**	96.29 (1.88)
PPV mean (SD)	91.47 (3.07)	92.10 (3.17)	90.78 (1.95)	92.34 (2.16)	95.25 (2.38)	**95.99 (2.26)**	97.71 (2.15)
NPV mean (SD)	87.17 (6.46)	**87.68 (6.65)**	75.12 (5.09)	79.78 (5.28)	86.68 (6.90)	87.56 (6.88)	93.70 (5.39)
Sensitivity mean (SD)	94.54 (3.64)	**94.65 (3.75)**	88.53 (3.38)	90.78 (3.31)	93.70 (4.28)	94.03 (4.24)	97.06 (2.78)
Specificity mean (SD)	78.94 (8.77)	80.55 (8.90)	78.88 (5.23)	82.23 (5.76)	88.80 (6.03)	**90.55 (5.72)**	94.52 (5.37)
F-score mean (SD)	92.87 (1.21)	93.24 (1.25)	89.59 (1.71)	91.49 (1.50)	94.37 (1.89)	**94.90 (1.82)**	97.34 (1.38)
Scenario 2							
ACC mean (SD)	79.03 (4.68)	79.71 (4.88)	71.93 (4.35)	74.76 (4.55)	81.79 (5.62)	**82.24 (5.27)**	90.45 (4.82)
PPV mean (SD)	86.59 (4.76)	86.97 (5.00)	84.66 (3.60)	86.11 (4.19)	**90.64 (5.18)**	90.56 (5.33)	94.37 (4.80)
NPV mean (SD)	64.76 (9.15)	66.18 (9.46)	51.51 (5.89)	55.52 (6.62)	68.61 (11.08)	**69.38 (10.42)**	84.63 (11.40)
Sensitivity mean (SD)	84.06 (8.20)	**84.74 (8.48)**	74.03 (5.63)	77.16 (5.95)	83.65 (10.09)	84.50 (9.45)	92.51 (7.18)
Specificity mean (SD)	66.72 (15.36)	67.38 (16.19)	66.79 (9.37)	68.89 (11.39)	**77.21 (15.06)**	76.70 (16.01)	85.42 (13.33)
F-score mean (SD)	84.91 (4.08)	85.41 (4.29)	78.85 (3.72)	81.20 (3.75)	86.42 (5.48)	**86.88 (5.07)**	93.14 (3.76)
Scenario 3							
ACC mean (SD)	83.68 (2.26)	83.98 (2.29)	79.55 (2.84)	81.49 (2.68)	**86.94 (2.81)**	86.83 (2.39)	91.94 (2.33)
PPV mean (SD)	86.71 (3.52)	86.62 (3.72)	87.94 (2.74)	88.44 (2.99)	**91.85 (3.64)**	91.09 (3.77)	95.36 (2.97)
NPV mean (SD)	76.81 (7.44)	78.42 (7.88)	62.95 (5.13)	66.93 (5.49)	77.71 (8.41)	**78.71 (8.11)**	85.25 (6.87)
Sensitivity mean (SD)	91.48 (4.82)	**92.16 (4.88)**	82.83 (4.17)	85.40 (4.28)	90.00 (5.38)	90.77 (5.11)	93.42 (3.91)
Specificity mean (SD)	64.24 (12.09)	63.61 (12.88)	71.38 (7.99)	71.76 (9.04)	**79.34 (11.18)**	77.02 (11.80)	88.25 (8.20)
F-score mean (SD)	88.86 (1.66)	89.12 (1.63)	85.21 (2.26)	86.78 (2.08)	90.72 (2.18)	**90.74 (1.83)**	94.28 (1.74)

3.4 Application to AREDS

3.4.1 Data description

In this study, we analyzed a total of 806 participants with moderate-to-severe AMD (AREDS AMD categories 2, 3, or 4) from the placebo group and antioxidants and zinc combination group (also known as AREDS formula or supplements) of the AREDS, who were free of late-AMD in at least one eye at enrollment. Table 3.4.1 summarized patients characteristics. Participants were randomized corresponding to AMD categories, and thus they did not differ on age, sex, smoking, and baseline AMD severity score between the two treatment arms.

We used the first eye progression time to late-AMD as the outcome of interest and the survival probability at 8 years (median overall survival time) as the treatment efficacy measure. In terms of the potential variables, we considered baseline characteristics including age at enrollment, smoking status, sex, education and baseline AMD severity scale, and 686 SNPs including 46 SNPs identified to be associated with treatment effect from Section 2.2 (Wei et al., 2020a) and 640 SNPs identified to be associated with AMD progression with $p < 10^{-5}$ and MAF > 0.05 (Yan et al., 2018). We first examined the overlap of the two cohorts regarding the confounders to check the violation of unconfoundedness assumption. Random forest classification (Liaw and Wiener, 2002) on the treatment assignment was used to estimate the propensity score (based on 4-fold cross validation). Figure 3.4.1 showed large overlap between the two groups, indicating the unconfoundedness assumption is valid in the AREDS dataset.

3.4.2 CATE estimation

Similar to the simulation studies, we applied the following methods to estimate CATE: R-T, R-X, B-T, B-X, D-T and D-X. In this analysis, we considered 3 random splits of data. For each split, 4-fold cross-validation was used to construct CATE estimates. For each method, individuals with CATE estimates ≥ 0 were labeled as RT (recommended for taking the treatment, i.e., AREDS supplements); otherwise, they were labeled RC (recommended

Table 3.4.1: Characteristics of the AREDS participants with AMD category 2, 3,or 4

Number of subjects	All ($n = 806$)	Placebo ($n = 391$)	Antioxidants and Zinc ($n = 415$)	p-value*
Age				0.4905
Mean (SD)	68.77 (5.05)	68.90 (5.17)	68.66 (4.93)	
Median (Range)	68.60 (55.30-81.00)	68.50 (55.30-81.00)	68.70 (55.50-79.50)	
Sex (n, %)				0.8236
Female	466 (57.82)	224 (57.29)	242 (58.31)	
Male	340 (42.18)	167 (42.71)	173 (41.69)	
Smoking (n, %)				0.6877
Never Smoked	393 (48.76)	194 (49.62)	199 (47.95)	
Former/Current Smoker	413 (51.24)	197 (50.38)	216 (52.05)	
AREDS AMD categories (n, %)				0.5474
2	312 (38.71)	158 (40.41)	154 (37.11)	
3	457 (56.70)	214 (54.73)	243 (58.55)	
4	37 (4.59)	19 (4.86)	18 (4.34)	
Baseline AREDS AMD severity score				0.6303
Mean (SD)	4.09 (2.06)	4.13 (2.06)	4.06 (2.07)	
Median (Range)	4 (1-8)	4.00 (1-8)	4 (1-8)	

$^\star p$-value is based on two-sample t-test or Chi-square test for continuous or categorical variables.

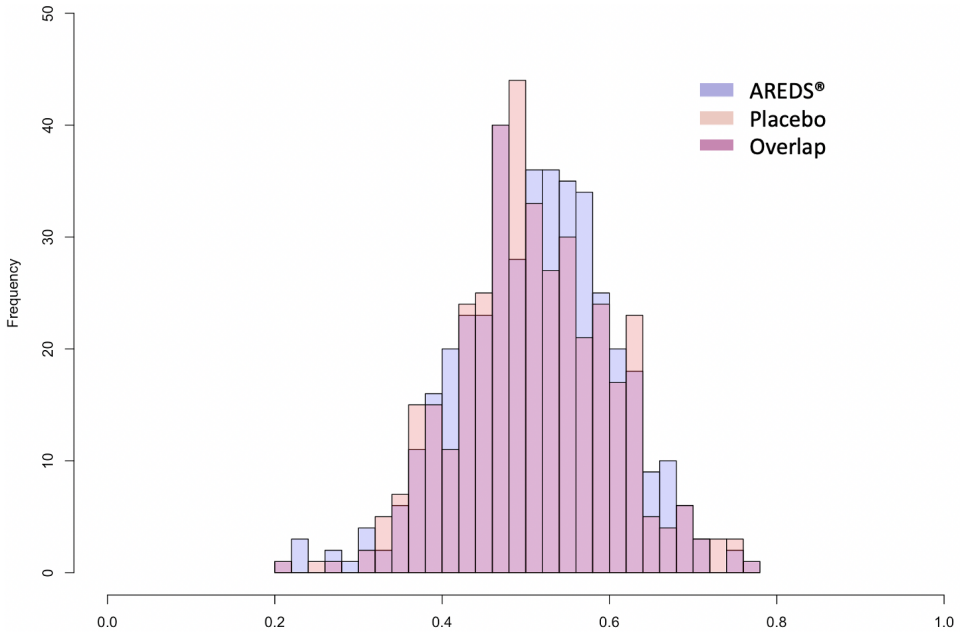

Figure 3.4.1: Distribution of estimated propensity score.

for taking control, i.e., placebo).

Under each method and each split of data, we compared the mean treatment effects within RT and RC by taking the difference of survival probabilities at 8 years between two treatment arms estimated through Kaplan-Meier (KM). Figure 3.4.2 visualized the mean treatment effect in RT cohorts defined by CATE estimate from each method, and the overall averaged treatment effect. In all splits of data, it showed positive treatment effect under recommendations of each method. Particularly, participants showed largest mean treatment effect under recommendations of D-T and D-X in all splits of data. Figure 3.4.3 showed mean treatment effect in RC cohorts from each method. Similarly, participants showed largest mean treatment effect under recommendations of D-T and D-X for placebo. Note that the treatment effect in RC cohorts is expected to be negative, since these participants should benefit more from taking the placebo as compared to taking the AREDS supplements.

In addition, we calculated the difference of survival probabilities at 8 years by using KM estimates between those recommended for the treatment and actually taking the treatment

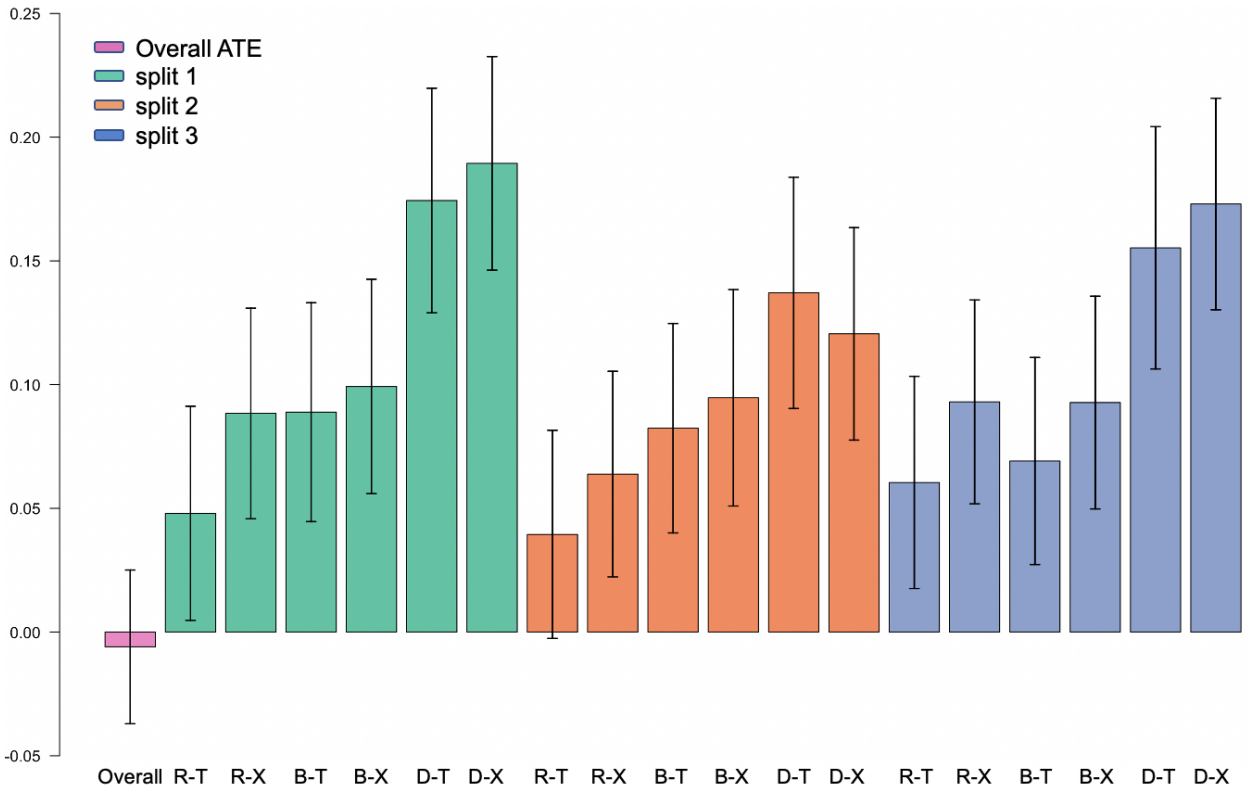

Figure 3.4.2: The mean treatment effect of participants recommended for treatment.

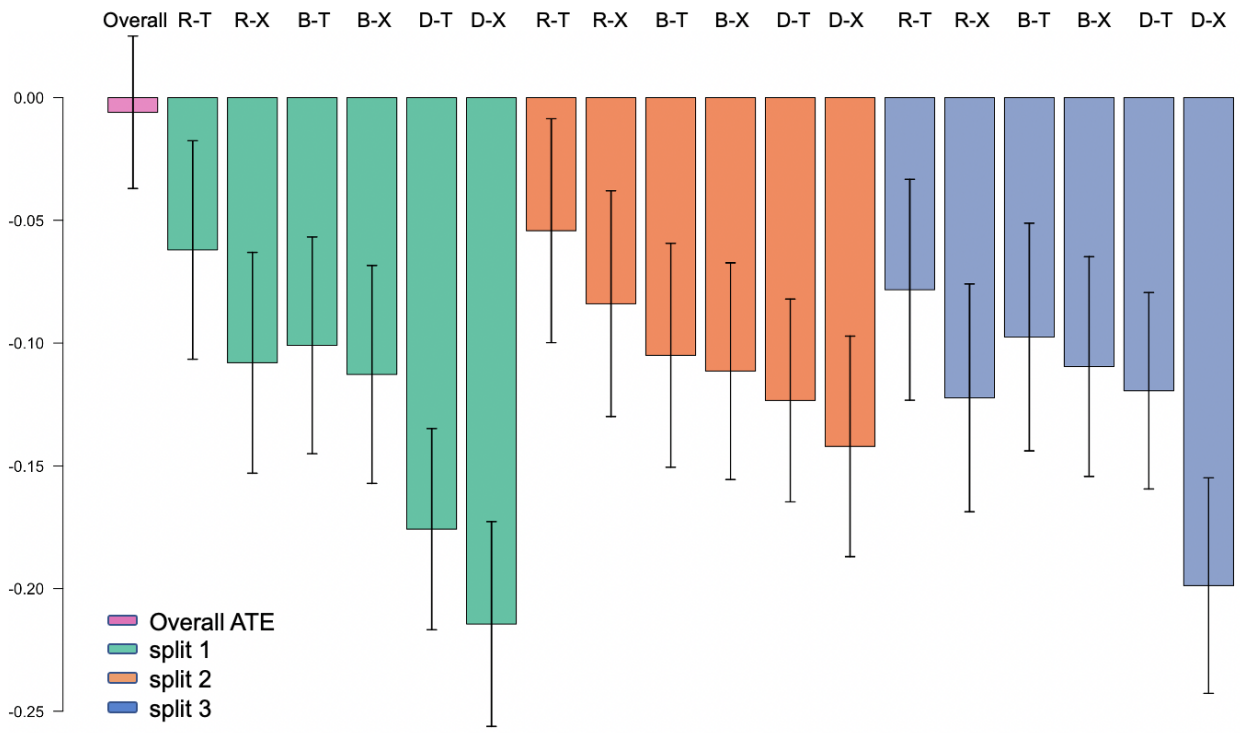

Figure 3.4.3: The mean treatment effect of participants recommended for placebo.

(RT & Taking treatment) and all individuals in the treatment group; and the difference of the survival probabilities at 8 years between those recommended for the placebo and actually taking the placebo (RC & Taking Placebo) and all individuals in the placebo group, respectively. From Figure 3.4.4, the estimated survival probability is higher in 'RT & Taking treatment' group than that in treatment group for all methods and all splits of data. Similar findings are observed by comparing those in 'RC & Taking Placebo' and those in placebo group. For 'RT & Taking Treatment' and 'RC & Taking Placebo', we can assume that the survival outcome of these subgroups are expected to be the survival if people were treated following the individual treatment recommendation. In that case, the survival probabilities among these people should be higher as compared to all individuals within the same treatment arm. Especially, in comparing survival probabilities between 'RT & Taking treatment' vs treatment group, D-X shows the largest differences, while in comparing survival probabilities between 'RC & Taking placebo' vs placebo group, both D-T and D-X show larger differences.

These analyses show that D-T and D-X perform the best among all of the methods we applied. These two methods recommend about 40% to 50% participants for treatment in each split of data with about 25% in the overlap across 3 splits: D-T recommends 394, 366, 388 out of 806 participants for treatment with 237 participants in the overlap of 3 splits of data; D-X recommends 414, 414, 418 out of 806 participants for treatment with 264 participants in the overlap of 3 splits of data.

3.4.3 Identification of important variables

Based on the analysis in the previous section, we chose to use CATE estimate from D-X to generate a treatment recommendation rule for patients. In the next step, we applied Boruta algorithm within each split of data to identify important variables in constructing the treatment recommendation rule (Kursa and Rudnicki, 2010). Confirmed variables were extracted from three splits: we identified 40 variables in split 1 with 39 SNPs consistent with those that were identified in Section 2.2 (denote as CE4-survival SNPs); 33 confirmed variables in split 2 with 30 SNPs from CE4-survival SNPs; 51 confirmed variables in split

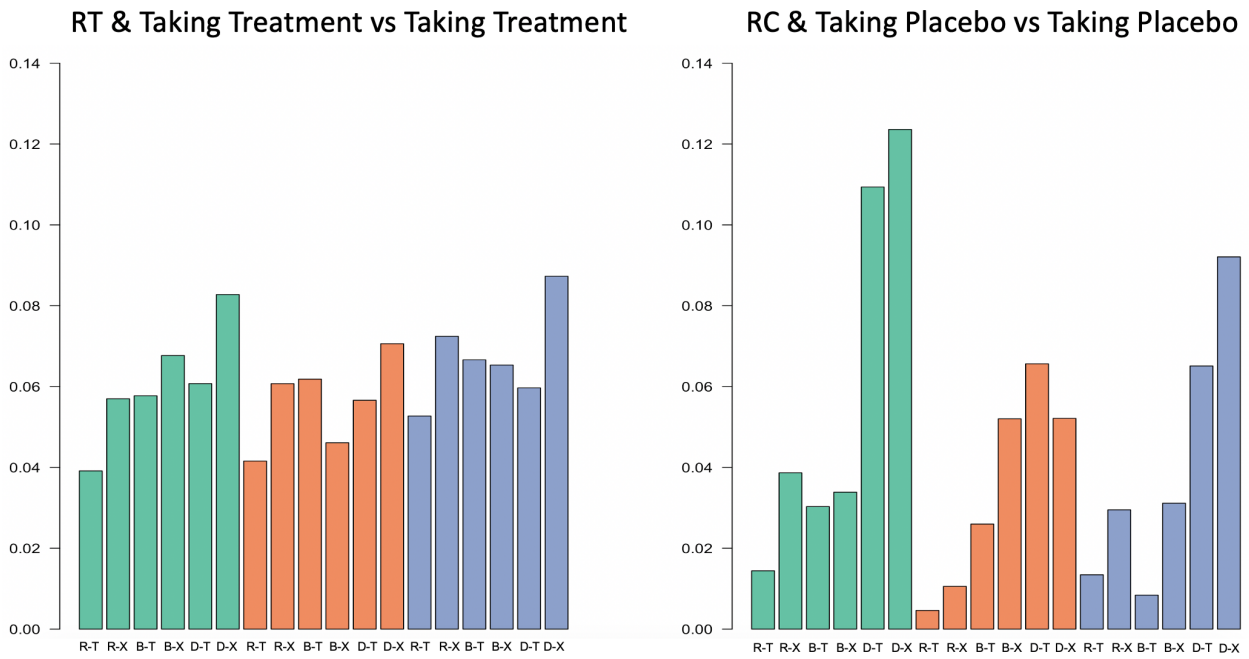

Figure 3.4.4: Difference of survival probabilities at 8 years by using KM estimates between those recommended for the treatment and actually taking the treatment (RT & Taking treatment) and those in the treatment group; and the difference of the survival probabilities at 8 years between those recommended for the placebo and actually taking the placebo (RC & Taking placebo) and those in the placebo group.

3 with 43 SNPs from CE4-survival SNPs. Across all identified SNPs from three splits, 23 SNPs were present to be "confirmed" important variables for all three splits and they were all from CE4-survival SNPs.

The SNPs we identified were mainly from chromosome 10, chromosome 19, and chromosome 14, which are the three regions mentioned in Section 2.2. We selected one SNP from each gene region and generated genotype distributions between RC and RT in 3 splits of data (Table 3.4.2). Participants who were recommended for placebo were less likely to have *aa* (less than the prevalence of *aa* in the overall population in 3 splits of data and in 3 selected SNPs); whereas participants who were recommended for the combination treatment were more likely to have *Aa* or *aa* (more than the prevalence of *Aa* and *aa* in 3 splits of data and in 3 selected SNPs). We used χ^2 test to see the difference of genotype distributions between RC and RT, and it is significant for all three splits of data and three selected SNPs.

Figure 3.4.5 shows the mean of ITE estimates with standard errors estimates from D-X in data split 1 and the estimated difference of survival probabilities at 8 years from Weibull model by genotype groups. The trends are consistent in data split 2 and 3 so we omitted the results. In the selected SNPs, mean ITE estimates were negative in *AA* group (-0.1 to 0) whereas positive in *Aa* and *aa* groups, with *aa* having largest mean of ITE estimates (>0.1). We observed similar trend from fitting Weibull regressions with treatment by corresponding SNP interaction, adjusting for age at enrollment, smoking status, and baseline severity score. The estimated difference of survival probabilities at 8 years in each genotype group was computed based on the parameters estimated from Weibull model. In summary, patients carrying at least one copy of minor allele in these SNPs are more likely to benefit from taking the AREDS supplements.

3.5 Conclusion and Discussion

In this work, we implemented the meta-algorithms on right-censored time-to-event outcomes using RSF, BAFT and DNNSurv as base learners, to estimate the CATE and provide a best treatment recommendation for patients between two treatment options to prolong

Table 3.4.2: Genotype distributions by the recommended treatment using D-X across 3 splits of data

			Split 1		**Split 2**		**Split 3**	
	Genotype	**Overall**	**RC**	**RT**	**RC**	**RT**	**RC**	**RT**
			n=392)	**(n=414)**	**(n=392)**	**(n=414)**	**(n=388)**	**(n=418)**
rs1245576 (*SPOCK2*, CHR10)	AA	266 (33.00%)	189 (48.21%)	77 (18.60%)	182 (46.43%)	84 (20.29%)	176 (45.36%)	90 (21.53%)
	Aa	408 (50.62%)	177 (45.15%)	231 (55.80%)	184 (46.94%)	224 (54.11%)	179 (46.13%)	229 (54.78%)
	aa	132 (16.37%)	26 (6.63%)	106 (25.60%)	26 (6.63%)	106 (25.60%)	33 (8.51%)	99 (23.68%)
rs8109218 (*C19orf44-CALR3*, CHR19)	AA	388 (48.14%)	238 (60.71%)	150 (36.23%)	244 (62.24%)	144 (34.78%)	228 (58.76%)	160 (38.28%)
	Aa	333 (41.32%)	134 (34.18%)	199 (48.07%)	134 (34.18%)	199 (48.07%)	136 (35.05%)	197 (47.13%)
	Aa	85 (10.55%)	20 (5.10%)	65 (15.70%)	14 (3.57%)	71 (17.15%)	24 (6.19%)	61 (14.59%)
rs147106198 (*ESRRB-VASH1*, CHR14)	AA	387 (48.01%)	237 (60.46%)	150 (36.23%)	232 (59.18%)	155 (37.44%)	243 (62.63%)	144 (34.45%)
	Aa	343 (42.56%)	139 (35.46%)	204 (49.28%)	142 (36.22%)	201 (48.55%)	136 (35.05%)	207 (49.52%)
	aa	76 (9.43%)	16 (4.08%)	60 (14.49%)	18 (4.59%)	58 (14.01%)	9 (2.32%)	67 (16.03%)

⋆All tests between RC and RT based on Chi-squared tests are statistically significant at the significance level of 0.05.

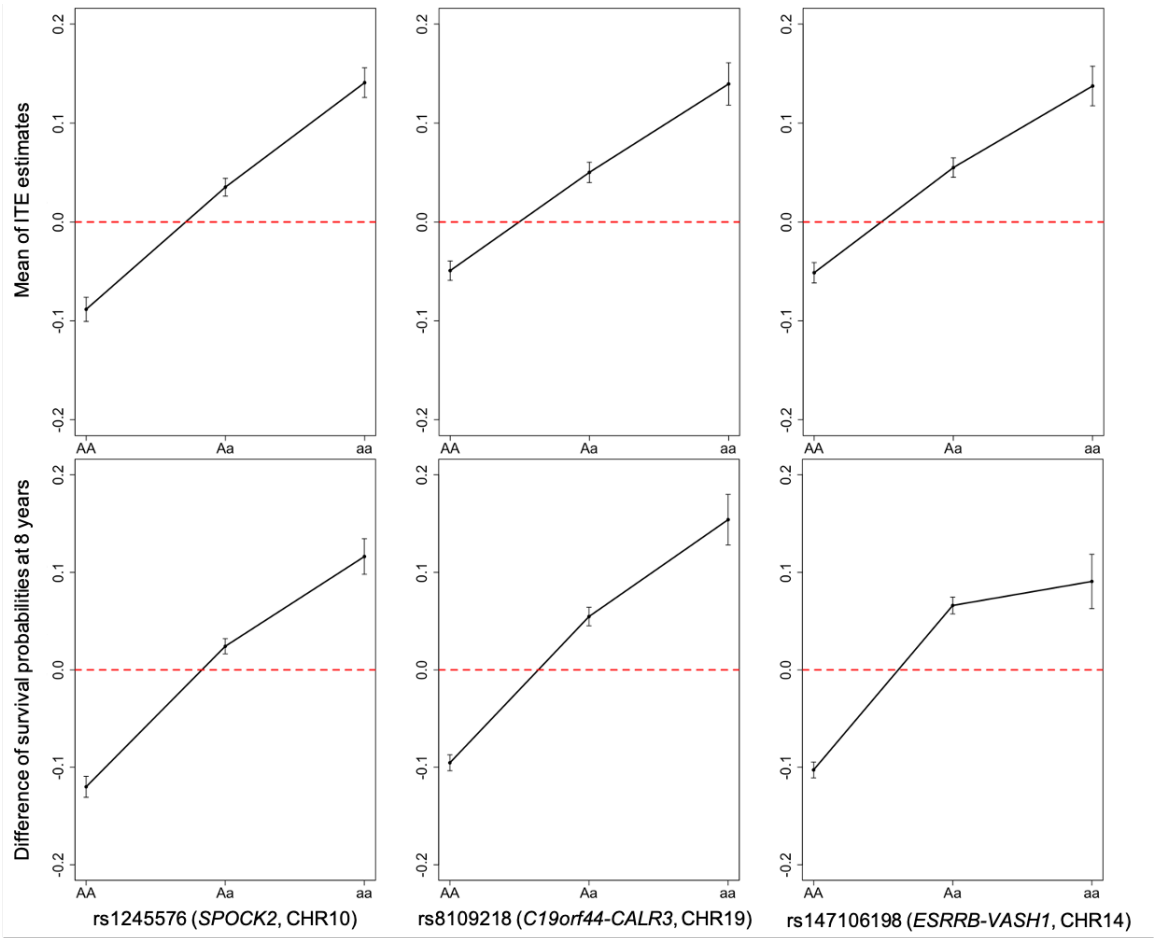

Figure 3.4.5: Treatment effect by genotype groups. The plots in the upper panel show the mean of ITE estimates from data split 1 using D-X by genotype groups. The plots in the lower panel show the difference of survival probabilities between the combination treatment and placebo groups at 8 years using Weibull regressions by genotype groups.

the time to late AMD. The performance of these methods was evaluated using intensive simulation studies. We observed that DNNSurv-based learners outperform other compared methods with smaller biases and RMSEs. RSF-based models generally have comparable biases with DNNSurv-based models, but the RMSEs from these models are relatively large. We applied these meta-learners on AREDS data to identify patients who can benefit from the AREDS supplements. T- and X-learner with DNNSurv model were used to identify the group of patients with greatest enhanced treatment effects. Lastly, Boruta algorithm was applied to identify important variables that contribute to constructing the treatment recommendation rule. Several SNPs were confirmed to be important variables in distinguishing patients with enhanced treatment effects. All of these SNPs are from regions that we know to be associated with differential treatment efficacy of using AREDS supplements in delaying the AMD progression from previous study. The trend of mean estimated individual treatment effects across genotype groups is consistent with the estimated group treatment effects from Weibull regressions.

Our study provides practical information which would be helpful to implement the meta-algorithm using the machine learning methods to estimate HTE and develop individual treatment recommendation. First, the simulation results suggest a limitation of using DNNSurv-based model. It requires a sizable cohort to train multiple tuning hyperparameters, and if the sample size is not large enough, the performance can be unstable. Second, under unbalanced design, the X-learner does not have dramatically improved performance as compared to the T-learner. The reason for that is we are now using the estimated survival probability in the second stage to construct imputed ITEs, where the original X-learner uses the observed value to serve as the most trustful values and gives it a larger weight. In our case, since there is no observed value for survival probability, by using the estimated values, the imputed treatment efficacy can be biased when the model is misspecified, and thus the performance of X-learner may not be better than the T-learner. To improve X-learner's performance, one can consider other approaches. For example, the observed progression status can be used as the true survival probability at certain time point to substitute the estimated values, and the censored data points needs to be imputed carefully.

4.0 Future work

In Chapter 2.2, we have developed a multiple testing based model to identify and infer subgroups with beneficial differential treatment efficacy using time-to-event outcomes. While AMD is a bilateral disease, the correlation between two eyes of the same individual needs to be carefully addressed, as mentioned in Chapter 2.3. In the future, I could extend the current approach to bivariate time-to-event data using copula models. Additionaly, the current approach is based on fully parametric AFT models. I am also interested in extending the method to semi-parametric models like CoxPH model, or semi-parametric AFT models, to have more flexibility. In Chapter 3, we implemented the X-learner with time-to-event data by using the predicted efficacy measure to substitute the original observed values from continuous and binary outcomes. However, this approach does not guarantee the gains of using X-learner since the model to predict efficacy measure may still be mis-specified and lead to biased estimates. I hope to modify the X-learner by using the inverse-weighted probability estimates of survival status at certain time point to substitute the observed value in the second step of X-learner. These new methods may facilitate precision medicine and subgroup identification.

Appendix Delta method for estimating the variance-covariance matrix of CE4 contrasts with time-to-event outcome

We demonstrate the use of Delta method to estimate the variance-covariance (var-cov) matrix of CE4 contrasts with time-to-event outcome. It includes two steps. We first used Delta method for implicit random variable (Benichou and Gail, 1989) to construct var-cov matrix for the following parameters:

$$(\log r_0, \log r_1, \log r_2, \nu^{Rx}_{12,\tau}, \nu^{C}_{12,\tau}, \nu^{Rx}_{01,\tau}, \nu^{C}_{01,\tau}).$$

Then the second step is to use Delta method for explicit random variable to estimate the var-cov of CE4 contrasts which are built from these parameters:

$$\begin{aligned}
\log \kappa_{(1,2):0} &= \log(\frac{r_{12}}{r_0}) = \log r_{12} - \log r_0 = \log \frac{\nu^{Rx}_{12,\tau}}{\nu^{C}_{12,\tau}} - \log r_0, \\
\log \kappa_{1:0} &= \log(\frac{r_1}{r_0}) = \log r_1 - \log r_0, \\
\log \kappa_{2:(0,1)} &= \log(\frac{r_2}{r_{01}}) = \log r_2 - \log r_{01} = \log r_2 - \frac{\nu^{C}_{01,\tau}}{\nu^{C}_{01,\tau}}, \\
\log \kappa_{2:1} &= \log(\frac{r_2}{r_1}) = \log r_2 - \log r_1. \qquad \text{(A.0.1)}
\end{aligned}$$

Starting from the following AFT model (equation (1)):

$$\begin{aligned}
\log T = \beta_0 + \beta_1 Trt + \beta_2 I(M=1) + \beta_3 I(M=2) + \\
\beta_4 Trt \times I(M=1) + \beta_5 Trt \times I(M=2) + \boldsymbol{\beta_6 X} + \sigma W, \qquad \text{(A.0.2)}
\end{aligned}$$

we further assume W follows Weibull distribution as demonstrated throughout the manuscript.

Denote $\exp(\beta_0)$ by λ, $\exp(-\frac{\beta_i}{\sigma})$ by θ_i, where $i = 1, \ldots, 5$ and $\exp(-\frac{\boldsymbol{\beta_6 X}}{\sigma})$ by θ_6. Let $A = (\log r_0, \log r_1, \log r_2, \nu^{Rx}_{12,\tau}, \nu^{C}_{12,\tau}, \nu^{Rx}_{01,\tau}, \nu^{C}_{01,\tau})$, the parameters we are interested in, and $B = (\lambda, \sigma, \theta_1, \ldots, \theta_6)$. By fitting a Weibull model, point estimates and var-cov matrix of B (denote by Σ) can be obtained. Benichou and Gail (Benichou and Gail, 1989) showed that if $G = (g_1, \ldots, g_p)$ is a vector of p functions of A and B, let J denote the $p \times p$ matrix with elements $\frac{\partial g}{\partial x}$, and let H denote the $p \times k$ matrix with elements $\frac{\partial g}{\partial y}$, where k is the number

of parameters in B ($k = 8$ in our setting), then the var-cov for the derived variates A is $J^{-1}H\Sigma H^{'}(J^{-1})^{'}$.

Step 1: Calculate the var-cov for A. Define $p_0 = \frac{P_{AA}}{P_{AA}+P_{Aa}}$ and $p_1 = \frac{P_{Aa}}{P_{Aa}+P_{aa}}$, where P_{AA}, P_{Aa}, P_{aa} represent the prevalence of genotype AA, Aa, and aa. The G function is defined as:

$$
\begin{aligned}
g_1 &= -\sigma \log \theta_1 - \log r_0, \\
g_2 &= -\sigma \log (\theta_1\theta_4) - \log r_1, \\
g_3 &= -\sigma \log (\theta_1\theta_5) - \log r_2, \\
g_4 &= p_1 \exp(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}) + (1-p_1)\exp(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}) - \tau, \\
g_5 &= p_1 \exp(-\theta_2\theta_6(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}) + (1-p_1)\exp(-\theta_3\theta_6(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}) - \tau, \\
g_6 &= p_0 \exp(-\theta_1\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}) + (1-p_0)\exp(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}) - \tau, \\
g_7 &= p_0 \exp(-\theta_6(\frac{\nu^{C}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}) + (1-p_0)\exp(-\theta_2\theta_6(\frac{\nu^{C}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}) - \tau.
\end{aligned}
$$

Then the J matrix is a diagonal matrix with $J_{11} = J_{22} = J_{33} = -1$,

$$
\begin{aligned}
J_{44} &= \frac{\partial g_4}{\partial \nu^{Rx}_{12,\tau}} = -\frac{\theta_1\theta_2\theta_4\theta_6 p_1}{\sigma\lambda}(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_1\theta_2\theta_4\theta_6)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- \frac{\theta_1\theta_3\theta_5\theta_6(1-p_1)}{\sigma\lambda}(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}), \\
J_{55} &= \frac{\partial g_5}{\partial \nu^{C}_{12,\tau}} = -\frac{\theta_2\theta_6 p_1}{\sigma\lambda}(\frac{\nu^{C}_{12,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_2\theta_6)(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- \frac{\theta_3\theta_6(1-p_1)}{\sigma\lambda}(\frac{\nu^{C}_{12,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_3\theta_6(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}), \\
J_{66} &= \frac{\partial g_6}{\partial \nu^{Rx}_{01,\tau}} = -\frac{\theta_1\theta_6 p_0}{\sigma\lambda}(\frac{\nu^{01}_{12,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_1\theta_6)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- \frac{\theta_1\theta_2\theta_4\theta_6(1-p_0)}{\sigma\lambda}(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}), \\
J_{77} &= \frac{\partial g_7}{\partial \nu^{C}_{01,\tau}} = -\frac{\theta_6 p_0}{\sigma\lambda}(\frac{\nu^{C}_{01,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_6)(\frac{\nu^{C}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- \frac{\theta_2\theta_6(1-p_0)}{\sigma\lambda}(\frac{\nu^{C}_{01,\tau}}{\lambda})^{(\frac{1}{\sigma}-1)}\exp(-\theta_2\theta_6(\frac{\nu^{C}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}).
\end{aligned}
$$

H matrix is a 7×8 matrix with the unit defined by $\frac{\partial g}{\partial Y}$.

$$
\begin{aligned}
H_{11} &= \frac{\partial g_1}{\partial \lambda} = 0, H_{12} = \frac{\partial g_1}{\partial \sigma} = -\log\theta_1, H_{13} = \frac{\partial g_1}{\partial \theta_1} = -\frac{\sigma}{\theta_1}, \\
H_{14} &= \cdots = H_{18} = 0, \\
H_{21} &= \frac{\partial g_2}{\partial \lambda} = 0, H_{22} = \frac{\partial g_2}{\partial \sigma} = -\log(\theta_1\theta_4), H_{23} = \frac{\partial g_2}{\partial \theta_1} = -\frac{\sigma}{\theta_1}, \\
H_{24} &= H_{25} = 0, H_{26} = \frac{\partial g_2}{\partial \theta_4} = -\frac{\sigma}{\theta_4}, H_{27} = H_{28} = 0, \\
H_{31} &= \frac{\partial g_3}{\partial \lambda} = 0, H_{32} = \frac{\partial g_3}{\partial \sigma} = -\log(\theta_1\theta_5), H_{33} = \frac{\partial g_3}{\partial \theta_1} = -\frac{\sigma}{\theta_1}, \\
H_{34} &= H_{35} = H_{36} = 0, H_{37} = \frac{\partial g_3}{\partial \theta_5} = -\frac{\sigma}{\theta_5}, H_{38} = 0, \\
H_{41} &= \frac{\partial g_4}{\partial \lambda} = \frac{p_1\theta_1\theta_2\theta_4\theta_6}{\lambda}\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_1)\theta_1\theta_3\theta_5\theta_6}{\lambda}\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{42} &= \frac{\partial g_4}{\partial \sigma} = \frac{p_1\theta_1\theta_2\theta_4\theta_6}{\sigma^2}\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln(\frac{\nu^{Rx}_{12,\tau}}{\lambda})(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_1)\theta_1\theta_3\theta_5\theta_6}{\sigma^2}\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln(\frac{\nu^{Rx}_{12,\tau}}{\lambda})(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{43} &= \frac{\partial g_4}{\partial \theta_1} = -p_1\theta_2\theta_4\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_1)\theta_3\theta_5\theta_6\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{44} &= \frac{\partial g_4}{\partial \theta_2} = -p_1\theta_1\theta_4\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{45} &= \frac{\partial g_4}{\partial \theta_3} = -(1-p_1)\theta_1\theta_5\theta_6\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{46} &= \frac{\partial g_4}{\partial \theta_4} = -p_1\theta_1\theta_2\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{47} &= \frac{\partial g_4}{\partial \theta_5} = -(1-p_1)\theta_1\theta_3\theta_6\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{48} &= \frac{\partial g_4}{\partial \theta_6} = -p_1\theta_1\theta_2\theta_4\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_1)\theta_1\theta_3\theta_5\exp\left(-\theta_1\theta_3\theta_5\theta_6(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{51} &= \frac{\partial g_5}{\partial \lambda} = \frac{p_1\theta_2\theta_6}{\lambda}\exp\left(-\theta_2\theta_6(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_1)\theta_3\theta_6}{\lambda}\exp\left(-\theta_3\theta_6(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{C}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}},
\end{aligned}
$$

$$
\begin{aligned}
H_{52} &= \frac{\partial g_5}{\partial \sigma} = \frac{p_1\theta_2\theta_6}{\sigma^2}\exp\left(-\theta_2\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln\left(\frac{\nu^C_{12,\tau}}{\lambda}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_1)\theta_3\theta_6}{\sigma^2}\exp\left(-\theta_3\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln\left(\frac{\nu^C_{12,\tau}}{\lambda}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{53} &= \frac{\partial g_5}{\partial \theta_1} = 0, \\
H_{54} &= \frac{\partial g_5}{\partial \theta_2} = -p_1\theta_6\exp\left(-\theta_2\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
H_{55} &= \frac{\partial g_5}{\partial \theta_3} = -(1-p_1)\theta_6\exp\left(-\theta_3\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{56} &= \frac{\partial g_5}{\partial \theta_4} = H_{57} = \frac{\partial g_5}{\partial \theta_5} = 0, \\
H_{58} &= \frac{\partial g_5}{\partial \theta_6} = -p_1\theta_2\exp\left(-\theta_2\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_1)\theta_3\exp\left(-\theta_3\theta_6(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{12,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{61} &= \frac{\partial g_6}{\partial \lambda} = \frac{p_0\theta_1\theta_6}{\lambda}\exp\left(-\theta_1\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_0)\theta_1\theta_2\theta_4\theta_6}{\lambda}\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{62} &= \frac{\partial g_6}{\partial \sigma} = \frac{p_0\theta_1\theta_6}{\sigma^2}\exp\left(-\theta_1\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln\left(\frac{\nu^{Rx}_{01,\tau}}{\lambda}\right)(\frac{\nu^{01}_{12,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_0)\theta_1\theta_2\theta_4\theta_6}{\sigma^2}\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln\left(\frac{\nu^{Rx}_{01,\tau}}{\lambda}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{63} &= \frac{\partial g_6}{\partial \theta_1} = -p_0\theta_6\exp\left(-\theta_1\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_0)\theta_2\theta_4\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{64} &= \frac{\partial g_6}{\partial \theta_2} = -(1-p_0)\theta_1\theta_4\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{65} &= 0, \\
H_{66} &= \frac{\partial g_6}{\partial \theta_4} = -(1-p_0)\theta_1\theta_2\theta_6\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{67} &= 0, \\
H_{68} &= \frac{\partial g_6}{\partial \theta_6} = -p_0\theta_1\exp\left(-\theta_1\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_0)\theta_1\theta_2\theta_4\exp\left(-\theta_1\theta_2\theta_4\theta_6(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^{Rx}_{01,\tau}}{\lambda})^{\frac{1}{\sigma}},
\end{aligned}
$$

$$
\begin{aligned}
H_{71} &= \frac{\partial g_7}{\partial \lambda} = \frac{p_0\theta_6}{\lambda}\exp\left(-\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_0)\theta_2\theta_6}{\lambda}\exp\left(-\theta_2\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\frac{1}{\sigma}(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{72} &= \frac{\partial g_7}{\partial \sigma} = \frac{p_0\theta_6}{\sigma^2}\exp\left(-\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln(\frac{\nu^C_{01,\tau}}{\lambda})(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&+ \frac{(1-p_0)\theta_2\theta_6}{\sigma^2}\exp\left(-\theta_2\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)\ln(\frac{\nu^C_{01,\tau}}{\lambda})(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}, \\
H_{73} &= \frac{\partial g_7}{\partial \theta_1} = 0, \\
H_{74} &= \frac{\partial g_7}{\partial \theta_2} = -(1-p_0)\theta_6\exp\left(-\theta_2\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
H_{75} &= \frac{\partial g_7}{\partial \theta_3} = H_{76} = \frac{\partial g_7}{\partial \theta_4} = H_{77} = \frac{\partial g_7}{\partial \theta_5} = 0, \\
H_{78} &= \frac{\partial g_7}{\partial \theta_6} = -p_0\exp\left(-\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}} \\
&- (1-p_0)\theta_2\exp\left(-\theta_2\theta_6(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}\right)(\frac{\nu^C_{01,\tau}}{\lambda})^{\frac{1}{\sigma}}.
\end{aligned}
$$

With J and H derived from these procedures, Σ obtained from Weibull model, the var-cov for $A = (\log r_0, \log r_1, \log r_2, \nu^{Rx}_{12,\tau}, \nu^C_{12,\tau}, \nu^{Rx}_{01,\tau}, \nu^C_{01,\tau})$ is computed as $J^{-1}H\Sigma H'(J^{-1})'$.

Step 2: Calculate var-cov for CE4 contrast. This step is based on Delta method for explicit random variable and the matrix can be easily obtained in R using function *deltamethod* from package {msm}.

Bibliography

Age-Related Eye Disease Study Research Group (1999). The age-related eye disease study (AREDS): design implications. AREDS report no. 1. *Controlled Clinical Trials*, 20(6):573–600.

Assel, M. J., Li, F., Wang, Y., Allen, A. S., Baggerly, K. A., and Vickers, A. J. (2018). Genetic polymorphisms of *CFH* and *ARMS2* do not predict response to antioxidants and zinc in patients with age-related macular degeneration. *Ophthalmology*, 125(3):391–397.

Barros, A. and Hirakata, V. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*, 3(21).

Benichou, J. and Gail, M. H. (1989). A delta method for implicitly defined random variables. *Am. Stat.*, 43(1):41–44.

Cascella, R., Strafella, C., Caputo, V., Errichiello, V., Zampatti, S., Milano, F., Potenza, S., Mauriello, S., Novelli, G., Ricci, F., Cusumano, A., and Giardina, E. (2018). Towards the application of precision medicine in age-related macular degeneration. *Progress in Retinal and Eye Research*, 63:132–146.

Chakravarthy, U., Wong, T. Y., Fletcher, A., Piault, E., Evans, C., Zlateva, G., Buggage, R., Pleil, A., and Mitchell, P. (2010). Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis. *BMC Ophthalmology*, 10(31).

Chen, W., Qian, L., Shi, J., and Franklin, M. (2018). Comparing performance between logbinomial and robust poisson regression models for estimating risk ratios under model misspecification. *BMC Medical Research Methodology*, 18(63).

Chew, E. Y., Clemons, T., SanGiovanni, J. P., Danis, R., Domalpally, A., McBee, W., Sperduto, R., and Ferris, F. L. (2012). The age-related eye disease study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology*, 119(11):2282–2289.

Chew, E. Y., Clemons, T. E., Agrón, E., Sperduto, R. D., Sangiovanni, J. P., Kurinij, N., Davis, M. D., and Age-Related Eye Disease Study Research Group (2013). Long-term effects of vitamins c and e, β-carotene, and zinc on age-related macular degeneration: AREDS report no. 35. *Ophthalmology*, 120(8):1604–1611.

Chew, E. Y., Klein, M. L., Clemons, T. E., Agrón, E., and Abecasis, G. R. (2015). Genetic testing in persons with age-related macular degeneration and the use of the AREDS supplements: to test or not to test? *Ophthalmology*, 122(1):212–215.

Ciampi, A., Negassa, A., and Lou, Z. (1995). Tree-structured prediction for censored survival data and the cox model. *Journal of Clinical Epidemiology*, 48:675–689.

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

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

Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. (2021). *Handbook of Multiple Comparisons*. Boca Raton: Chapman & Hall/CRC.

Cui, Y., Kosorok, M. R., Wager, S., and Zhu, R. (2020). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv:2001.09887*.

Ding, Y., Li, Y. G., Liu, Y., Ruberg, S. J., and Hsu, J. C. (2018). Confident inference for snp effects on treatment efficacy. *The Annals of Applied Statistics*, 12(3):1727–1748.

Ding, Y., Lin, H.-M., and Hsu, J. C. (2016). Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine*, 35(10):1580–1594.

Ding, Y., Liu, Y., Yan, Q., Fritsche, L. G., Cook, R. J., Clemons, T., Ratnapriya, R., Klein, M. L., Abecasis, G. R., Swaroop, A., Chew, E. Y., Weeks, D. E., and Chen, W. (2017). Bivariate analysis of age-related macular degeneration progression using genetic risk scores. *Genetics*, 206:119–133.

Flaherty, K. T., Yasothan, U., and Kirkpatrick, P. (2011). Vemurafenib. *Nature Review Drug Discovery*, 10(11):811–812.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011a). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011b). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880.

Fritsche, L. G., Chen, W., and et al., M. S. (2013). Seven new loci associated with age-related macular degeneration. *Nature Genetics*, 45(4):433–439.

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., Burdon, K. P., Hebbring, S. J., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, 48:134–143.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2020). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-1.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2020). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.

Group, A.-R. E. D. S. (2001). A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins c and e, beta carotene, and zinc for age-related macular degeneration and vision loss: Areds report no. 8. *Archives of Ophthalmology*, 119(10):1417–1436.

Gumbel, E. J. (2012). *Statistics of extremes*. Courier Corporation.

Halekoh, U., Højsgaard, S., and Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference, What If*. Boca Raton: Chapman & Hall/CRC.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240.

Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38(6):226–231.

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

Hsu, J. C. and Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association*, 94:468–482.

Huitfeldt, A., Stensrud, M. J., and Suzuki, E. (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology*, 16(1):226–231.

Ishwaran, H. and Kogalur, U. (2007). Random survival forests for r. *R News*, 7(2):25–31.

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

Kang, S., Wenbin, L., and Rui, S. (2017). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 36:4646–4659.

Klein, M. L., Francis, P. J., Rosner, B., Reynolds, R., Hamon, S. C., Schultz, D. W., Ott, J., and Seddon, J. M. (2008). *CFH* and *LOC387715/ARMS2* genotypes and treatment with antioxidants and zinc for age-related macular degeneration. *Ophthalmology*, 115(6):1019–1025.

Kunzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2018). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, 36:1–13.

Lettre, G., Lange, C., and Hirschhorn, J. (2007). Genetic model testing and sta- tistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31(4):358–362.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lin, H.-M., Xu, H., Ding, Y., and Hsu, J. C. (2019). Correct and logical inference on efficacy in subgroups and their mixture for binary outcomes. *Biometrical Journal*, 61(1):8–26.

Lipkovich, I., Dmitrienko, A., and D'Agostino Sr., R. B. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621.

Liu, Y., Tang, S.-Y., Man, M., Li, Y. G., Ruberg, S. J., Kaizar, E., and Hsu, J. C. (2016). Thresholding of a continuous companion diagnostic test confident of efficacy in targeted population. *Statistics in Biopharmaceutical Research*, 8(3):325–333.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34:1818–1833.

Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2017). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 0(0):1–11.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15(3):231–239.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Davidson, N. E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P. A., Swain, S. M., Pisansky, T. M., Fehrenbacher, L., Kutteh, L. A., Vogel, V. G., Visscher, D. W., Yothers, G., Jenkins, R. B., Brown, A. M., Dakhil, S. R., Mamounas, E. P., Lingle, W. L., Klein, P. M., and Ingle, J. N. (2005). Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. *The New England Journal of Medicine*, 353(16):1673–1684.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(4):688–701.

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

SanGiovanni, J. P., Chen, J., Sapieha, P., Aderman, C. M., Stahl, A., Clemons, T. E., Chew, E. Y., , and Smith, L. E. H. (2013). DNA sequence variants in *PPARGC1A*, a gene encoding a coactivator of the ω-3 *LCPUFA* sensing *PPAR-RXR* transcription complex, are associated with nv amd and amd-associated loci in genes of complement and *VEGF* signaling pathways. *PLOS One*, 8(1):e53155.

Seddon, J., Silver, R., and Rosner, B. (2016). Response to AREDS supplements according to genetic factors: survival analysis approach using the eye as the unit of analysis. *British Journal of Ophthalmology*, 100(12):1731–1737.

Seddon, J. M., Francis, P. J., George, S., Schultz, D. W., Rosner, B., and Klein, M. L. (2007). Association of *CFH Y402H* and *LOC387715 A69S* with progression of age-related macular degeneration. *The Journal of the American Medical Association*, 297(16):1793–1800.

Seibold, H., Zeileis, A., and Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics*, 12(1):45–63.

Shaw, A. T., Yeap, B. Y., Solomon, B. J., Riely, G. J., Gainor, J., Engelman, J. A., Shapiro, G. I., Costa, D. B., Ou, S. H., Butaney, M., Salgia, R., Maki, R. G., Varella-Garcia, M., Doebele, R. C., Bang, Y. J., Kulig, K., Selaru, P., Tang, Y., Wilner, K. D., Kwak, E. L., Clark, J. W., Iafrate, A. J., and Camidge, D. R. (2011). Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. *Lancet Oncology*, 12(11):1004–1012.

Splawa-Neyman, J., Dabrowska, D., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, 5(4):465–472.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1):2.

Sun, T. and Ding, Y. (2019). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, 22:315–330.

Sun, T., Wei, Y., Chen, W., and Ding, Y. (2020). Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine*, 39:4605–4620.

Tabib, S. and Larocque, D. (2020). Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics*, 36(2):629–636.

Touloumis, A. (2019). Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. R package version 1.7.0.

Vavvas, D. G., Small, K. W., Awh, C. C., Zanke, B. W., Tibshirani, R. J., and Kustra, R. (2018). *CFH* and *ARMS2* genetic risk determines progression to neovascular age-related macular degeneration after antioxidant and zinc supplementation. *Proceedings of the National Academy of Sciences of the United States of America*, 115(4):E696–E704.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228–1242.

Wakusawa, R., Abeb, T., Satoa, H., Yoshidaa, M., Kunikataa, H., Satoc, Y., and Nishidaa, K. (2008). Expression of vasohibin, an antiangiogenic factor, in human choroidal neovascular membranes. *American Journal of Ophthalmology*, 146(2):235–243.

Warnes, G. (2019). *genetics: Population Genetics*. R package version 1.3.8.1.2.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879.

Wei, Y., Hsu, J. C., Chen, W., Chew, E. Y., and Ding, Y. (2020a). A simultaneous inference procedure to identify subgroups from rcts with survival outcomes: Application to analysis of amd progression studies.

Wei, Y., Liu, Y., Sun, T., Chen, W., and Ding, Y. (2020b). Gene-based association analysis for bivariate time-to-event data through functional regression with copula models. *Biometrics*, 76:619–629.

Williamson, T., Eliasziw, M., and Fick, G. H. (2013). Log-binomial models: exploring failed convergence. *Emerging Themes in Epidemiology*, 10.

Wu, R.-f., Zheng, M., and Yu, W. (2016). Subgroup analysis with time-to-event data under a logistic-cox mixture model. *Scandinavian journal of statistics*, 43:863–878.

Xu, Y., Yu, M., Zhao, Y., Li, Q., Wang, S., and Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, 71(3):645–653.

Yan, Q., Ding, Y., Liu, Y., Sun, T., Fritsche, L. G., Clemons, T., Ratnapriya, R., Klein, M. L., Cook, R. J., Liu, Y., Fan, R., Wei, L., Abecasis, G. R., Swaroop, A., Chew, E. Y., Group, A. R., Weeks, D. E., and Chen, W. (2018). Genome-wide analysis of disease progression in age-related macular degeneration. *Human Molecular Genetics*, 27(5):929–940.

Yelland, L. N., Salter, A. B., and Ryan, P. (2011). Performance of the modified possion regression approach for estimating relative risks from clustered prospective data. *American Journal of Epidemiology*, 174(8):984–992.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zeng, S., Hernandez, J., and Mullins, R. F. (2012). Effects of antioxidant components of AREDS vitamins and zinc ions on endothelial cell activation: implications for macular degeneration. *Investigative Ophthalmology & Visual Science*, 53(2):1041–1047.

Zhang, B., Tsiatis, A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, P., Ma, J., Chen, X., and Shentu, Y. (2020). A nonparametric method for value function guided subgroup identification via gradient tree boosting for censored survival data. *Statistics in Medicine*, 39(28):4133–4146.

Zhao, Y., Zheng, D., Rush, A., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(449):1106–1118.

Zhu, J. and Gallego, B. (2020). Targeted estimation of heterogeneous treatment effect in observational survival analysis. *Journal of Biomedical Informatics*, 107:103474.

Zou, G. (2003). A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7):702–706.