Use and Evaluation of Statistical Methods for Personalized Medicine in Oncology

by

Pin Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Associate Professor Matthew J. Schipper, Co-Chair
Professor Jeremy M.G. Taylor, Co-Chair
Associate Professor Phil S. Boonstra
Professor Theodore S. Lawrence

Pin Li

pinli@umich.edu

ORCID iD: 0000-0003-4508-3762

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my co-advisor, Prof. Jeremy Taylor for his guidance and advice. His vision, diligence, motivation and patience have deeply inspired me. It was a great privilege and honor to work and study under his guidance.

I would like to express my heartiest gratitude to my co-advisor, Prof. Matthew Schipper for his guidance and support. He has provided me with invaluable research opportunities that help me build confidence as a researcher. I would also like to thank him for his encouragement, reliability and great sense of humor.

Many thanks to my dissertation committee member, Prof. Phil Boonstra and Prof. Theodore S. Lawrence for their helpful suggestions in the methodology and application in my dissertation research.

I would like to thank my supervisors and collaborators in my GSRA projects, including Prof. Jill Becker from Department of Psychology, Prof. Ananda Sen and Prof. Wen Ye from Department of Biostatistics, Prof. Michelle Mierzwa, Prof. Yue Cao, Prof. Michelle Kim, Prof. Daniel Spratt and Dr. Laila Gharzai from Department of Radiation Oncology.

I would like to thank all the faculty members and staffs in the Biostatistics Department for establishing a nurturing environment for future scientists. My sincere thanks also go to Prof. Veera Baladandayuthapani and the TaBaBoo research group for their many helpful comments and support. I would also like to thank my fellows and friends in the Biostatistics Department.

Last but not least, I am extremely grateful to my parents for their love, caring and sacrifices for educating and preparing me for my future. I want to thank my husband Dr. Wei Zhao, who has always been a great and patient listener to my struggles and a firm supporter of the decisions I've made. They all kept me going and I could not make this far without their trust and support.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

The goal of personalized medicine is to give the right treatment to the right patient at the right dose using all we know about the patient. With the increasing availability of biomarkers and prediction models, there is the potential for individualized treatment based on patient specific factors. There are many statistical challenges associated with achieving this goal. One is how to develop and assess good predictions models. Another is how to define a criteria for an optimal treatment when there are multiple outcomes and then how to analyze available data to determine the optimal treatment for each future patient.

In Chapter II, we consider the assessment of prediction models using data with missing biomarker values. We propose inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimates of the area under the ROC curve (AUC) and Brier Score to handle the missing data. AIPW is a double-robust method that is robust to the misspecification of either a model for the missingness mechanism or a model for the distribution of the missing variable. We evaluated the performance of IPW and AIPW in comparison with multiple imputation (MI) in simulation studies under missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) scenarios. We illustrate these methods using an example from prostate cancer.

In Chapters III and IV we consider the setting where there is an existing dataset of patients treated with heterogeneous doses and including binary efficacy and toxicity outcomes and patient factors such as clinical features and biomarkers. The goal is to

analyze the data to estimate an optimal dose for each (future) patient based on their clinical features and biomarkers.

In Chapter III, we propose an optimal individualized dose finding rule by maximizing utility functions for individual patients while limiting the rate of toxicity. The utility is defined as a weighted combination of efficacy and toxicity probabilities. We model the binary efficacy and toxicity outcomes using logistic regression with dose, biomarkers and dose-biomarker interactions. To incorporate the large number of potential parameters, we use the LASSO method. We additionally constrain the dose effect to be non-negative for both efficacy and toxicity for all patients. The proposed methods are illustrated using a dataset of patients with lung cancer treated with radiation therapy.

In Chapter IV, we extend the approach of Chapter III and propose to use flexible machine learning methods such as random forests and Gaussian processes to build models for efficacy and toxicity depending on the dose and biomarkers. In addition, we allow for dependence between efficacy and toxicity. A copula is used to model the joint distribution of the two outcomes and the estimates are constrained to have non-decreasing dose-efficacy and dose-toxicity relationships. Numerical utilities are assigned to each potential outcome pair, which allow the improvement in the utility due to a change in efficacy to depend on the level of toxicity. For each patient, the optimal dose is chosen to maximize the utility function or the posterior mean of the utility function. We further adjust the utility function with more constraints to incorporate clinical requirements, and consider the uncertainty in the estimation of the utility function in the optimal dose selection. The various models and methods are evaluated in a simulation study and illustrated using data from a lung cancer study.

CHAPTER I

Introduction

Traditionally, standards of care have been built based on knowledge from clinical studies and evidence-based medicine and given in medical practice. With the progress in basic science, scientists are developing and using diagnostic tests based on genetics or other molecular mechanisms to better predict patients' responses to targeted therapy. The use of biomarkers provides the potential to drive therapeutic decision making and tailor medical care. Personalized medicine has been focusing on optimizing the assignment of therapy by treating patients with the right drug at the right dose at the right time.

The personalized medicine approach can replace the traditional "one size fits all" medicine and move from a reactive to a proactive discipline that is predictive, personalized, preventive and participatory [17]. However, the field of personalized medicine raises many challenges that need to be overcome, such as integration of diverse data from the various hierarchical levels of biological information, difficulties identifying and validating molecular markers, and ultimately create effective predictive and actionable models. Besides technical challenges, there are also societal challenges including education and communication with patients, physicians, and the health-care community about the power of personalized medicine to accelerate the application of it.

The statistical challenges associated with personalized medicine are to provide effective tools for construction and evaluation of evidence-based personalized intervention strategies. Multiple regression has been widely used for model building to identify predictive biomarkers with data from randomized clinical trials (RCT). In the absence of randomization, the propensity score can be used to reweight the observational data to attenuate selection bias. With a large set of biomarkers, variable selection methods are needed to detect predictive biomarkers from high-dimensional sources. More advanced methods are proposed to improve the robustness against misspecification of regression models, and well established classification methods and other popular machine learning techniques can alternatively be customized to avoid pre-specification of parametric models [30].

Because models are estimated conditionally given the observed data, they neglect to quantify the extent of uncertainty for future patients. Thus external validation and internal validation are needed, and cross-validation (CV) and bootstrap resampling techniques are commonly used for the later. To assess the model performance, the Brier score (BS) and receiver operating characteristic (ROC) curves are two common metrics used to evaluate existing prediction models of a binary outcome. The assessment of prediction models using data with missing covariate values is challenging. There are general methods for handling missing data of multiple imputation (MI) and inverse probability weighting [27, 41]. These methods make use of specific statistical models, which have to be selected. How to select these models is an important question. When using MI to develop a prediction model, Moons et. al. [33] showed that it is important to include the outcome variable within the selected imputation models. How to construct the statistical models in either MI or IPW methods to give valid results when assessing an exisiting prediction model is not known.

In Chapter II, we propose inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimates of the area under the ROC curve

(AUC) and BS to handle the missing data. AIPW is a double-robust method that is robust to the misspecification of either a model for the missingness mechanism or a model for the distribution of the missing variable [1]. An alternative approach uses multiple imputation (MI), which requires a model for the distribution of the missing variable. We evaluated the performance of IPW and AIPW in comparison with MI in simulation studies under missing complete at random (MCAR), missing at random (MAR), and missing not at random (MNAR) scenarios. When there are missing observations in the data, MI and IPW can be used to obtain unbiased estimates of BS and AUC if the imputation model for the missing variable or the model for the missingness is correctly specified. MI is more efficient than IPW. AIPW can improve the efficiency of IPW, and also achieves double robustness from misspecification of either the missingness model or the imputation model. The outcome variable should be included in the model for the missing variable under all scenarios, while it only needs to be included in missingness model if the missingness depends on the outcome. We illustrate these methods using an example from prostate cancer.

In early phase clinical trials of oncology, efficacy such as disease progression and toxicity such as side effects are used to evaluate the performance of a treatment. Different approaches can be used to measure the trade-off between potentially beneficial and potentially harmful outcomes for a patient, such as utility function [40], utility matrix [19] or utility contours [46]. There has been much work on how to choose the individualized optimal treatment when there is a choice between two treatments and there is a single outcome measure [12, 54, 53, 55]. There has been much less research on methods to find individualized treatment rule in the setting with more than two treatments and more than one outcome. We consider the setting where there is an existing dataset of patients treated with heterogeneous doses and including binary efficacy and toxicity outcomes and patient factors such as clinical features and biomarkers. The goal is to analyze the data to estimate an optimal dose for each

(future) patient based on their clinical features and biomarkers. With the increasing availability of validated biomarkers, there is the need to identify the patient specific factors that is related to the individualized dosing rule. In many settings, including oncology, increasing the dose of treatment usually results in both increased efficacy and toxicity, and how to incorporate this is challenging. The efficacy and toxicity outcomes could be correlated or independent given the dose, and both of them should be used to decide the optimal dose.

In Chapter III, we propose an optimal individualized dose finding rule by maximizing utility functions for individual patients while limiting the rate of toxicity [49]. The utility is defined as a weighted combination of efficacy and toxicity probabilities. This approach maximizes overall efficacy at a prespecified constraint on overall toxicity. We model the binary efficacy and toxicity outcomes using logistic regression with dose, biomarkers and dose-biomarker interactions. To incorporate the large number of potential biomarkers, we use the LASSO method [47]. We additionally constrain the dose effect to be non-negative for both efficacy and toxicity for all patients. Simulation studies show that the utility approach combined with any of the modeling methods can improve efficacy without increasing toxicity relative to fixed dosing. The proposed methods are illustrated using a dataset of patients with lung cancer treated with radiation therapy.

In Chapter IV, we extend the approaches in Chapter III by relaxing the assumption of independence of efficacy and toxicity in modeling. We propose to use flexible machine learning methods such as random forest [24] and Gaussian process [51] to build models for efficacy and toxicity depending on the dose and biomarkers. Copula [43] is used to model the joint distribution of the two outcomes and the estimates are constrained to have non-decreasing dose-efficacy and dose-toxicity relationships. Numerical utilities are assigned to each potential outcome pair, and these allow the improvement in the utility due to a change in efficacy to depend on the level of toxicity.

For each patient, the optimal dose is chosen to maximize the utility function or the posterior mean of the utility function. We further adjust the utility function with more constraints to incorporate clinical requirements, and consider the uncertainty in the estimation of the utility function in optimal dose selection. The various models and methods are evaluated in a simulation study, where we find that using Gaussian processes to model the probability of efficacy and toxicity and using the posterior mean of the utility function has good properties. We also compare the proposed random forest and Gaussian process methods with the parametric models in Chapter III to illustrate their pros and cons.

To conclude, the goal is to construct and evaluate models, which can accurately predict either the clinical outcome or the treatment effect for individual patients, and then optimize the outcome or treatment effect for new patients. This dissertation makes a contribution towards the model estimation and evaluation in oncology and proposes novel methodologies in optimal individualized dose finding.

CHAPTER II

Evaluation of Predictive Model Performance of an Existing Model in the Presence of Missing Data

2.1 Introduction

In clinical research, patient information such as clinical features, diagnostic tests and biomarkers are often used to help with diagnosis or to provide prognosis of a future outcome for a patient with disease. When the outcome of interest is binary, a typical prediction model will numerically combine the covariates, for example using a linear combination, to estimate the predicted probability of the binary outcome. The evaluation of an existing prediction model in a different populations is of considerable interest. If a model is to be transportable to other populations, it needs to be validated, which is usually thought of as meaning that it has similar and good performance in other populations. The performance of existing prediction models can be assessed using a variety of metrics, such as the Brier score to indicate accuracy of the probabilistic predictions, and area under the receiver operating characteristic (ROC) curve (or the concordance statistic) for discriminative ability [44]. Very often, covariate values will be missing for some patients. The assessment of prediction models in data with missing covariate values is a challenge. The context we are considering is that the existing model or models were developed on other datasets, which we call

the external data, and are already completely specified. We do not have access to the data used to develop these models. Rather, our goal is to assess the performance of the existing model in an available dataset, which we call the internal data, that has missing values for some covariates and we want to get valid and efficient estimates of the BS and the AUC.

In general there are two types of methods for estimation in the presence of missing data, one is based on multiple imputation (MI) and the other is based on inverse probability weighting (IPW). For MI, a model for the distribution of the missing variable, or variables, needs to be specified. For IPW method, a model for the probability of missingness needs to be specified, which is also called the weight model. For multiple imputation, M completed datasets are created and M model performance measures can be estimated from each of the completed dataset and then averaged [27]. Alternatively, an overall measure of model performance can be estimated directly from a simple completed dataset that includes the average of the M predictions for each missing value. As previously recommended [52], the former is preferred. The analysis of only the observations with complete data is frequently biased, and inverse probability weighting is a commonly used approach to correct their bias [41]. It is also used to adjust for unequal sampling fractions in sample surveys and causal inference [31]. Augmented inverse probability weighting (AIPW) has been proposed as an extension of IPW. It is a double-robust method that is robust to the misspecification of either a model for the missingness mechanism or a model for the distribution of the variables with missing values (but not both) [1]. AIPW generally results in improved efficiency compared to IPW, although this is not guaranteed to be the case..

When analyzing data with missing values an important consideration is the missingness mechanism, and the mechanism will impact the properties and merits of different methods. Missing complete at random (MCAR) is when the probability of any variable being missing for a subject does not depend on the value of any of the the

variables. Generally all methods will work under MCAR. Analysis of the complete cases will be unbiased, but are frequently quite inefficient compared to other methods, depending on the amount of missingness. Missing at random (MAR) is when the probability of being missing can depend on other covariates, but only those that are observed. In general MI, IPW and AIPW are vaild under MAR, if models are appropriately specified. Complete case analysis is frequently biased under MAR. Missing Not at Random (MNAR) is when the probability of missing depends on the value of variables that are fully observed, including the unobserved value of the variable itself. Generally all methods are biased under MNAR.

A basic question for all the above MI, IPW and AIPW methods is whether the observed data for the outcome variable should be included in the required imputation models or weight models. This is also related to how the covariate is missing, whether the missingness is completely at random, or depend on other covariates and/or the outcome, or the covariate itself. The argument in favor of including the outcome variable in these models is from the theoretical developments associated with missing data and multiple imputation. In general, it is well known that for inference about a quantity of interest it is necessary to include the outcome variable as one of the variables in the imputation model when developing a new prediction model. Omitting the outcome variable can lead to biased estimates [33]. In general notation, if Q is the quantity of interest, and $D = (D_{obs}, D_{mis})$ is the data where D_{obs} and D_{mis} denote the observed and missing data, then from a Bayesian perspective, inference about Q is based on its posterior distribution $P(Q|D_{obs})$. This posterior distribution can be written as $P(Q|D_{obs}) = \int P(Q|D_{obs}, D_{mis})P(D_{mis}|D_{obs})pD_{mis}$, and this applies whether Q is a simple parameter in a model or a more complex function i.e. such as the Brier Score or the AUC. This formula is the recipe for multiple imputation and motivates imputation of the missing data using $P(D_{mis}|D_{obs})$, followed by inference for Q using the complete data (D_{obs}, D_{mis}), and repeating these steps many times and averaging

them. Since in our setting the outcome variable is part of D_{obs}, it is clear that the outcome variable should be used as part of the imputation scheme. In practice, the general recommendation for MI is that the imputation model should include every variable that predicts the incomplete variable, and sometimes the imputation model can contain more variables than will be used in the final analysis [50].

The intuitive argument against including the outcome variable in the models used for imputation is the belief that there is some circularity. Since we are trying to evaluate how good a model is at predicting outcome, the thinking is that we don't want to use the outcome to help impute the missing covariates, because then we will make the model look better than it really is. However, Moons et al argued that imputation of missing values using all other information will not create information. It only makes use of the strength of associations between predictors and outcomes present in the complete cases, to enable valid analyses [33]. The additional intuitive argument against using the outcome variable is that the intended use of these models is in the situation where we want to make a prediction for a single patient and we only have covariates available and the outcome is not known. It is certainly a challenge of how to make a prediction if some of the covariates are missing, but this is a different situation than ours of evaluating a existing prediction model using a new dataset.

In this paper, we propose IPW and AIPW estimates of AUC and Brier score to handle the missing data and evaluate their prediction performance in comparison with MI by simulation. We focus on including the outcome or not in the weight models or imputation models. The missing mechanisms could be MCAR, MAR and MNAR. We consider a variety of existing prediction models including ones that are both consistent with and not consistent with the internal data distribution, and ones that depend on a subset of the covariates. An example from prostate cancer is used as an illustration of the proposed methods.

2.2 Methods

We consider the setting in which we have available an internal dataset of size N, consisting of binary outcome Y and p-dimensional vector of covariates X. Let $R_i = 1$ if there are no missing X values for subject i, else $R_i = 0$ if there are missing values. Assume there is an existing external model, that requires as input the variables X or a subset of the variables, and produces as output an estimate of the probability that $Y = 1$, denoted by $\hat{p}(Y = 1|X)$. We use notation I to denote distributions associated with the available or internal data, and E to denote the distributions associated with the external data that was used to build the existing model. Let $F_I(X)$ and $F_I(Y|X)$ denote the true probability distribution functions for the internal data. Thus $F_I(X)$ is the density of X if X is continuous and $F_I(Y = y|X = x) = P_I(Y = y|X = x)$. Let $F_E(X)$ and $F_E(Y|X)$ denote the true distributions for the external data. We would expect some of the X's to be correlated with each other.

The existing model $\hat{p}(Y = 1|X)$ is an approximation to $F_E(Y = 1|X)$, and it is usually a monotonic function of a weighted combination of covariates, denoted as $g(\beta X)$. The estimates of β could be good estimates if for example the external dataset is large and good methods were used, or they could be poor estimates if the external dataset is small or poor methods were used. From the internal dataset with sample size N that are sampled from $F_I(X)$ and $F_I(Y|X)$, we can calculate the Brier score and AUC. The BS is given by

$$BS = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{p_i})^2 \tag{2.1}$$

where $\hat{p_i}$ =P$(Y = 1|X_i)$ is obtained from the existing model.

The Area Under the Curve (AUC), which is equivalent to the Concordance-index

(C-index) for a binary outcome, is estimated using

$$AUC/C-index = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)} \tag{2.2}$$

An alternative way to estimate the AUC is to first estimate the ROC curve and then calculate the area under it. Let n_1 denote the number of cases, n_0 denote the number of controls, and $n_1 + n_0 = N$. Let X_1 denote the covariates in cases and X_0 denote the covariates in controls. The ROC curve depicts relative trade-offs between true positive rate (TPR) and false positive rate(FPR),

$$\begin{aligned} TPR(c) &= \Pr(\beta X_1 \geq c) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(\beta X_i \geq c) \\ FPR(c) &= \Pr(\beta X_0 \geq c) = \frac{1}{n_0}\sum_{j=n_1+1}^{N} I(\beta X_j \geq c) \\ ROC(c) &= TPR(FPR^{-1}(c)) \\ AUC &= \int_0^1 ROC(c)dc \end{aligned} \tag{2.3}$$

The integration of ROC to calculate the AUC is performed numerically. The quantities called BS and AUC given above are estimates of population quantities, which we call $TrueBrier_I(\hat{p})$ and $TrueAUC_I(\hat{p})$. Given the distribution $F_I(X)$ and $F_I(Y|X)$, for any existing formula $\hat{p}$ that provides a probability that Y=1 given X, the true Brier Score (BS) is defined as

$$TrueBrier_I(\hat{p}) = \sum_{Y=0}^{1}\int_X (Y-\hat{p})^2 F_I(Y|X)F_I(X)dX \tag{2.4}$$

For covariates in cases X_1 and controls X_0, denote their distributions as $F_I(X_1) = F_I(X|Y=1)$ and $F_I(X_0) = F_I(X|Y=0)$, respectively. Then the true AUC is

$$TrueAUC_I(\hat{p}) = \Pr(\beta X_1 > \beta X_0) = \int_{X_1}\int_{X_0} I(\beta X_1 > \beta X_0) F_I(X_1) F_I(X_0) dX_1 dX_0 \tag{2.5}$$

Equation 2.4 and 2.5 give the true values of BS and AUC for a fixed β. The goal is to get good estimates of these population quantities $\text{TrueAUC}_I(\hat{p})$ and $\text{TrueBrier}_I(\hat{p})$, using the available data in the internal dataset of size N. A good estimate is one that has small bias, low variability and is robust to model misspecification.

Also note from equation 2.4 that the true value depends on both $F_I(Y|X)$ and $F_I(X)$, and similarly for equation 2.5. This makes it clear that even if the existing prediction model for Y given X is correct for the internal distribution, it will not usually lead to the same AUC and BS because these depend on the X distribution as well. In practice it would seem likely that the internal and external distributions of the X's do differ.

In real data analysis with large sample size, missing data are a common occurrence. Suppose our dataset has missing values for some covariates of X, and the missingness may be MCAR, MAR or MNAR. The practical question we are trying to address is how to get a good estimate of $\text{TrueAUC}_I(\hat{p})$ and $\text{TrueBrier}_I(\hat{p})$ from the available dataset with missing covariates. The best conceivable estimates are the ones that would have been obtained using equations 2.1,2.2 and 2.3 if there had been no missing data.

2.2.1 Complete case analysis

Using only complete case (i.e $R_i = 1$) the simplest estimates are

$$BS_{CC} = \frac{\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i}{\sum_{i=1}^{N} R_i} \tag{2.6}$$

$$C - index_{CC} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i R_j}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j) R_i R_j} \tag{2.7}$$

For AUC,

$$\begin{aligned} TPR_{CC}(c) &= \frac{\sum_{i=1}^{n_1} I(\beta X_i \geq c) R_i}{\sum_{i=1}^{n_1} R_i} \\ FPR_{CC}(c) &= \frac{\sum_{j=1}^{n_0} I(\beta X_j \geq c) R_j}{\sum_{j=1}^{n_0} R_j} \end{aligned} \tag{2.8}$$

However, these estimates may be biased in MAR and MNAR settings and may lack efficiency in MCAR situations.

2.2.2 Multiple Imputation

When there is partially missing in X, we can do Multiple Imputation (MI) to impute the missing values based on the available data, then average the predicted BS and AUC from the multiple imputed datasets using Rubin's rule. The first step is to impute the missing values by drawing a value of X_{mis} from a model either for $F(X_{mis}|X_{obs}, Y)$ or for $F(X_{mis}|X_{obs})$, and then apply the external model on the imputed complete data to get the predictions of Y and calculate BS and AUC. The models used for imputation are typically linear regression for continuous X_{mis}, logistic

regression for binary X_{mis}, polytomous regression for unordered categorical X_{mis}, and proportional odds model for ordered categorical X_{mis}, although more complicated models could be used. After repeating the first step for M times (we use M=5), the average of the estimates of BS and AUC from the multiple imputed datasets gives the final single point estimate. When there is more than one covariate with missing values, a chained equation approach is used to impute the missing values sequentially [50]. The program `mice()` in `R` is used to implement the multiple imputations and the different models mentioned above can be built with options.

2.2.3 Inverse Probability Weighting

Inverse Probability Weighting (IPW) weights the complete cases in the calculation of the quantity of interest. The weight (W_i) is the inverse probability of the observation being complete ($R_i = 1$), i.e. $W_i = 1/\Pr(R_i = 1)$. We use logistic regression to build the model of either $\Pr(R_i = 1|X_i, Y_i)$ or $\Pr(R_i = 1|X_i)$ conditional on the fully observed covariates and outcome to get the estimates of the weight. Then

$$BS_{IPW} = \frac{\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i W_i}{\sum_{i=1}^{N} R_i W_i} \tag{2.9}$$

$$C - index_{IPW} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i W_i R_j W_j}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j) R_i W_i R_j W_j} \tag{2.10}$$

For AUC,

$$TPR_{IPW}(c) = \frac{\sum_{i=1}^{n_1} I(\beta X_i \geq c) R_i W_i}{\sum_{i=1}^{n_1} R_i W_i}$$
$$FPR_{IPW}(c) = \frac{\sum_{j=1}^{n_0} I(\beta X_j \geq c) R_j W_j}{\sum_{j=1}^{n_0} R_j W_j} \tag{2.11}$$

With the TPR_{IPW} and FPR_{IPW}, ROC_{IPW} and AUC_{IPW} can be calculated following (2.3).

2.2.4 Augmented Inverse Probability Weighting

The IPW method only uses the complete cases, and ignores the subjects with missing data. One way to improve it is to include information from subjects with missing data, which is called Augmented Inverse Probability Weighting (AIPW). For ease of notation we describe the method in the situation of only one covariate having missing values. In the Appendix we describe how to apply it when multiple covariates have missing values. First we build a model for the covariate with missing values on all the other covariates, i.e. $F(X_{mis}|X_{obs}, Y)$ or $F(X_{mis}|X_{obs})$, to get the predicted mean X^*_{mis}, which is $E(X_{mis}|X_{obs}, Y)$ or $E(X_{mis}|X_{obs})$. This is a single imputation of the mean and is different from multiple imputation which incorporates random variation. The X^*_{mis} is created for that variable for all subjects and is different from MI which only imputes missing values. Then applying the external model to the dataset with X replaced by $X^* = (X^*_{mis}, X_{obs})$ gives $\hat{p_i}^*$. Combining this model with a model for the weight, the AIPW estimator of the BS is

$$BS_{AIPW} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i W_i + (Y_i - \hat{p_i}^*)^2(1 - R_i W_i), \tag{2.12}$$

A subject with complete data has $R_i = 1$, and contributes $(Y_i - \hat{p}_i)^2 W_i + (Y_i - \hat{p_i}^*)^2(1 - W_i)$. A subject with missing values has $R_i = 0$ and contributes $(Y_i - \hat{p_i}^*)^2$. Because all the subjects with complete data or missing values are evaluated, the denominator is N.

For the C-index,

$$C{-}index_{AIPW} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)\big\{I(\beta X_i > \beta X_j) R_i W_i R_j W_j + I(\beta X_i^* > \beta X_j^*)(1 - R_i W_i R_j W_j)\big\}}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)} \tag{2.13}$$

A pair of cases and controls X_i, X_j that are both complete has $R_i = 1, R_j = 1$, and contributes $I(\beta X_i > \beta X_j) W_i W_j + I(\beta X_i^* > \beta X_j^*)(1 - W_i W_j)$. Otherwise, a pair of cases and controls that has missing value i.e $R_i = 0$ and/or $R_j = 0$ contributes $I(\beta X_i^* > \beta X_j^*)$.

For the area under the ROC curve method of calculating the AUC,

$$\begin{aligned} TPR_{AIPW}(c) &= \frac{1}{n_1}\sum_{i=1}^{n_1} I(\beta X_i \geq c) R_i W_i + I(\beta X_i^* \geq c)(1 - R_i W_i) \\ FPR_{AIPW}(c) &= \frac{1}{n_0}\sum_{j=1}^{n_0} I(\beta X_j \geq c) R_j W_j + I(\beta X_j^* \geq c)(1 - R_j W_j) \end{aligned} \tag{2.14}$$

A subject with complete data has $R_i = 1$, and contributes $I(\beta X_i \geq c) W_i + I(\beta X_i^* \geq c)(1 - W_i)$. A subject with missing value has $R_i = 0$ and contributes $I(\beta X_i^* \geq c)$. With the TPR_{AIPW} and FPR_{AIPW}, ROC_{AIPW} and AUC_{AIPW} can be calculated following (2.3).

2.2.5 Consistency of IPW and AIPW estimators

Considering the C-index using the IPW method. Let

$$U_{ij}(\theta, \gamma_1) = \theta I(Y_i > Y_j) R_i W_i R_j W_j - I(Y_i > Y_j) I(\beta X_i > \beta X_j) R_i W_i R_j W_j,$$

where W_i depend on weight model with parameters γ_1.

Let $U_N(\theta, \gamma_1) = 0.5N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \big[U_{ij}(\theta, \gamma_1) + U_{ji}(\theta, \gamma_1)\big]$, then it is straight forward to show that $C - index_{IPW}$ is the solution of $U_N(\theta, \gamma_1) = 0$. Let $U_E = E(U_N) = 0.5E\big[U_{ij}(\theta, \gamma_1) + U_{ji}(\theta, \gamma_1)\big]$.

Let γ_1^* be the large sample limit of $\hat{\gamma}_1$ using the weight model $\Pr(R = 1|X_{obs}, Y; \gamma_1)$. When the weight model is correctly specified, i.e. $\Pr(R = 1|X_{obs}, Y; \gamma_1^*) = \Pr(R = 1|X_{obs}, Y)$, then $E(R_i W_i R_j W_j) = 1$, and it is clear that $U_E(\theta, \gamma_1^*) = 0$. Because $U_N(\theta, \gamma_1)$ converges uniformly to $U_E(\theta, \gamma_1)$, $C - index_{IPW}$ is a consistent estimator.

The proof for AIPW estimators is similar. Here we mimic the proof in Long et. al. [29], and first demonstrate double robustness for a slightly different estimator, which we label $C - index_{AIPW*}$ with

$$C{-}index_{AIPW*} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I(Y_i > Y_j) \big\{ I(\beta X_i > \beta X_j) R_i W_i R_j W_j + E[I(\beta X_i > \beta X_j)](1 - R_i W_i R_j W_j) \big\}}{\sum_{i=1}^{N} \sum_{j=1}^{N} I(Y_i > Y_j)}$$

Let

$$V_{ij}(\theta, \gamma_1, \gamma_2) = \theta I(Y_i > Y_j) - I(Y_i > Y_j) \big\{ I(\beta X_i > \beta X_j) R_i W_i R_j W_j + E[I(\beta X_i > \beta X_j)](1 - R_i W_i R_j W_j) \big\}$$

where W_i depend on weight model ($\Pr(R = 1|X_{obs}, Y; \gamma_1)$) with parameters γ_1 and in $E[I(\beta X_i > \beta X_j)]$ the expectation is with respect to the distribution of the missing covariates and depend on the model ($F(X_{mis}|X_{obs}, Y; \gamma_2)$) with parameters γ_2.

Let $V_N(\theta, \gamma_1, \gamma_2) = 0.5N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \big[V_{ij}(\theta, \gamma_1, \gamma_2) + V_{ji}(\theta, \gamma_1, \gamma_2)\big]$, then it is straightforward to see that $C - index_{AIPW*}$ is the solution of $V_N(\theta, \gamma_1, \gamma_2) = 0$. Let $V_E = E(V_N) = 0.5E\big[V_{ij}(\theta, \gamma_1, \gamma_2) + V_{ji}(\theta, \gamma_1, \gamma_2)\big]$. It is easy to see that $V_N(\theta, \gamma_1, \gamma_2)$ converges uniformly to $V_E(\theta, \gamma_1, \gamma_2)$, thus the solution to $V_N(\theta, \gamma_1, \gamma_2) = 0$ converges to the solution of $V_E(\theta, \gamma_1, \gamma_2) = 0$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{obs}, Y; \gamma_1)$. When the weight model is correctly specified, i.e. $\Pr(R = 1|X_{obs}, Y; \gamma_1^*) = \Pr(R = 1|X_{obs}, Y)$, then $E(R_i W_i R_j W_j) = 1$.

Let γ_2^* be the probability limits of γ_2 using the model for the missing covariates $F(X_{mis}|X_{obs}, Y; \gamma_2)$. When the model is correctly specified, i.e., $E(X_{mis}|X_{obs}, Y; \gamma_2^*) = E(X_{mis}|X_{obs}, Y)$, then $E\big\{I(Y_i > Y_j)\{E[I(\beta X_i > \beta X_j)] - I(\beta X_i > \beta X_j)\}\big\} = 0$.

When either working model is correctly specified, it is clear that $V_E(\theta, \gamma_1, \gamma_2) = 0$, and that the θ that solves $V_E(\theta, \gamma_1, \gamma_2) = 0$ is the true AUC. Because V_N converges uniformly to V_E, $C - index_{AIPW*}$ is a consistent estimator.

The estimator we describe in section 2.4, $C - index_{AIPW}$ is an approximation to $C - index_{AIPW*}$, in which instead of calculating $E[I(\beta X_i > \beta X_j)]$ over the distribution, we propose to use $I(\beta X_i^* > \beta X_j^*)$.

The proof of consistency is similar for Brier score and is shown in the Appendix.

2.3 Simulation Studies

In this section, we present results of numerical studies to investigate the performance of the proposed method under different settings. We consider three covariates and denote them as X_1, X_2, X_3. We consider situations where the given external model is based on all of X_1, X_2 and X_3, and situations where it is only based on X_1 and X_2. The true distribution for the internal data, $F_I(Y|X)$, is defined as

$$logit(Pr(Y = 1)) = 0.25 + 0.7X_1 + 0.6X_2 - 0.5X_3$$

The internal data are sampled from above model. X_1, X_2, X_3 are sampled from $N(0, 1)$ and about 40% of X_1 is missing. The covariates can be independent, or correlated with $cor(X_1, X_3) = -0.5$. Four different external models are evaluated using the "internal" data; (M_1) the true model with X_1, X_2 and X_3; (M_2) the best model based on just X_1 and X_2; (M_3) a poor model based on X_1, X_2 and X_3 with wrong coefficients; and (M_4) an incorrect intercept model.

The simulation is conducted as follows:

(a) For M_1, we use the true coefficients, $M_1 = (0.25, 0.70, 0.60, -0.50)$. For M_2, we obtain the coefficients for the external model by generating a data set of 100000 observations from the true model, and fitting a logistic model based on X_1 and X_2. For independent covariates, $M_2 = (0.25, 0.67, 0.58, 0)$. For $cor(X_1, X_3) = -0.5, M_2 = (0.25, 0.91, 0.58, 0)$. It is noted that with independent covariates, the estimated coefficients are biased toward the null compared to the true model [35]. With correlated X_1, X_3 and X_3 is omitted, the estimates of the coefficients for X_1, X_2 are biased in opposite directions in the reduced model. For M_3, we obtain the coefficients by generating an external dataset with sample size 50. For independent covariates, $M_3 = (0.26, 0.66, 0.90, 0.39)$, and for correlated covariates, $M_3 = (0.53, -0.40, 0.88, -0.75)$. With such small sample size, the estimated coefficients are not close to the true values. For M_4, we set different prevalence's for the external data and internal data, and $M_4 = (1.00, 0.70, 0.60, -0.50)$.

(b) Based on the distributions $F_I(X), F_I(Y|X)$, get the true AUC and BS for each of M_1, M_2, M_3 and M_4 using their coefficients and equations 2.4 and 2.5. We label these as the true target values.

(c) Sample internal data with $N = 1000$, and evaluate the external models M_1, M_2, M_3, M_4 on the internal data. Use different methods to handle the miss-

Table 2.1: List of methods for comparison. * indicates methods for which the weight model is misspecified under MAR(X,Y). † indicates methods for which the imputation model is misspecified.

True target	true value based on internal data distribution
Full	data without missing
CC	complete cases analysis
IPW1*	weight model uses X
IPW2	weight model uses X & Y
MI1†	imputation model uses X
MI2	imputation model uses X & Y
AIPW1*†	weight model uses X, imputation model uses X
AIPW2†	weight model uses X & Y, imputation model uses X
AIPW3*	weight model uses X, imputation model uses X & Y
AIPW4	weight model uses X & Y, imputation model uses X & Y

ing covariates in the internal data to estimate AUC and BS, repeat 1000 times to get the mean and standard deviation, and compare with each other and with the true target value calculated in (b).

We consider four different missingness mechanisms. For MCAR, the missing of X_1 is random with probability 0.4, i.e., Pr(X_1 is missing)=0.4. For MAR(X_2, X_3), the missing of X_1 depends on other covariates X_2, X_3, Pr(X_1 is missing)= $expit(-0.5 + 2X_2 - 2X_3)$. For MAR($X_2, Y$), the missing of X_1 depends on both covariate X_2 and outcome Y, Pr(X_1 is missing)= $expit(-0.5 + 2X_2 + Y)$. For MNAR, the missing of X_1 depends on the value of X_1, Pr(X_1 is missing)= $expit(-0.5 + 3X_1)$.

As listed in Table 2.1, we compared the validation of external models on full internal data without missing (Full), on complete cases only (CC), IPW with the weight model excluding outcome Y (IPW1) or including outcome Y (IPW2), MI with the imputation model excluding outcome Y (MI1) or including outcome Y (MI2). When calculating AUC by AIPW, the two methods, which are based on the C-index and the area under the ROC curve respectively, gave almost identical results in terms of bias and efficiency, thus we show the results for the C-index using a weight

model that excludes the outcome Y (AIPW1, AIPW3) or includes the outcome Y (AIPW2, AIPW4) and using an imputation model that excludes the outcome Y (AIPW1, AIPW2) or includes the outcome Y (AIPW3, AIPW4). For the IPW and AIPW methods the weight models are regarded as mis-specified in the $MAR(X_2, Y)$ situation if they don't include Y, i.e. IPW1, AIPW1 and AIPW3, and all IPW and AIPW weight models are mis-specified in the MNAR situation.

In this simulation, `mice()` in `R` with linear regression using bootstrap is used to implement MI for the missing continuous covariates. `glm()` with logistic link was used to build weight models and `lm()` was used to calculate the predicted X_1^* in the AIPW method.

2.3.1 Simulation results

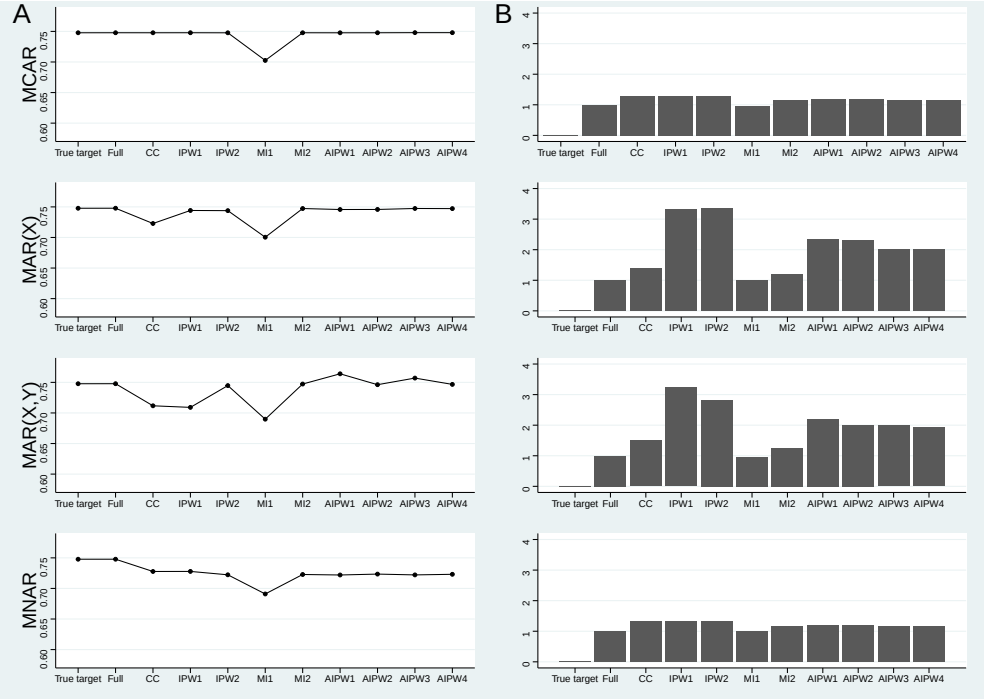

Figure 2.1: Simulation results of mean and relative SD of AUC for existing model M_1: correct model. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

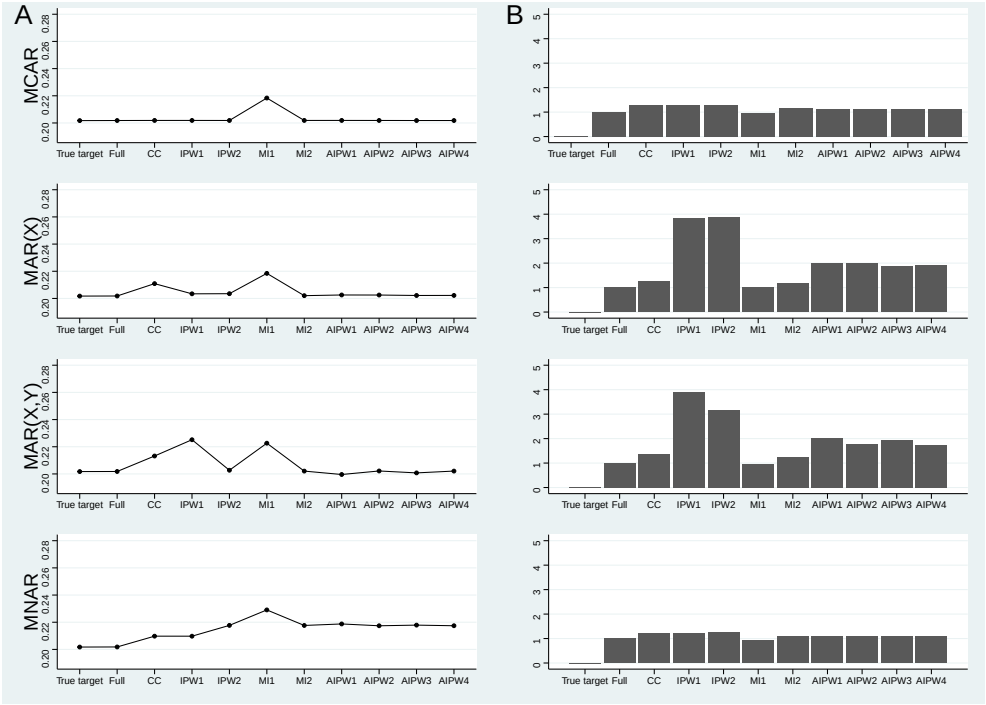

Figure 2.2: Simulation results of mean and relative SD of Brier score for existing model M_1: correct model. Column A denotes mean BS. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

Fig.2.1 and Fig.2.2 show the simulation results of AUC and BS for existing model M_1 with independent covariates under MCAR, MAR(X_2, X_3), MAR(X_2, Y), and MNAR(X_1). Column A shows the bias of the various methods. As expected the full data analysis does achieve the true target AUC and BS. However, the complete case analysis is unbiased only in the MCAR setting. MI with Y (MI2) is unbiased under MCAR and MAR, but without Y (MI1) the bias is more than 10% for both AUC and Brier score. All the IPW and AIPW methods are unbiased under MCAR and MAR(X_2, X_3), regardless of whether Y is included or not. Under MAR(X_2, Y) when Y is related to the missingness, the only unbiased IPW method (IPW2) is the one including Y, which indicates the importance of correct specification of the weight model. For AIPW2 and AIPW4, when the weight model includes Y, the results are unbiased. Without Y in the weight model, AIPW3 includes Y in the imputing model,

and the results are unbiased too. However, when both weight model and imputing model exclude Y, as in AIPW1, the results are biased, especially for AUC. For the double robustness of AIPW, as least one of the weight model and imputing model need to be correctly specified. Under MNAR for which the missingness depend on X_1, all the methods are biased.

Column B shows the relative SD of the methods comparing with full data estimation. As expected all values are equal to 1.0 or larger. The variance of IPW is always the largest, since it only weights the complete cases. The variance of AIPW is between IPW and MI, and is much smaller than IPW under MAR.

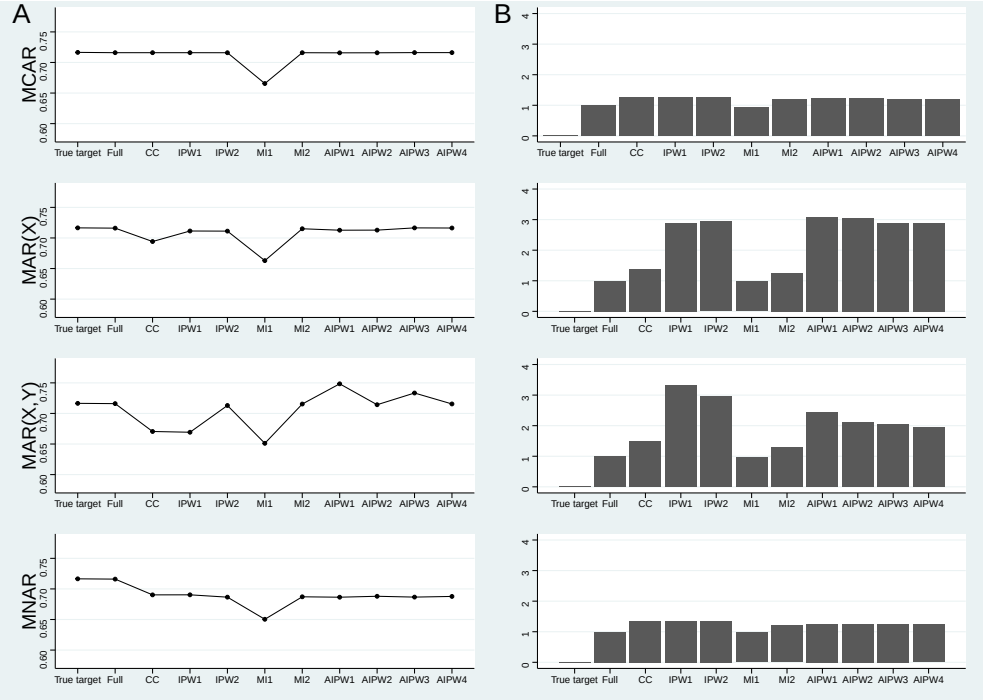

Figure 2.3: Simulation results of mean and relative SD of AUC for existing model M_2: best model based on just X_1, X_2. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

For the model M_2 with omitted covariate X_3, under all scenarios, the reduced model M_2 has lower AUC and higher Brier score compared with true target values for model M_1. This is to be expected since omitting an important covariate will

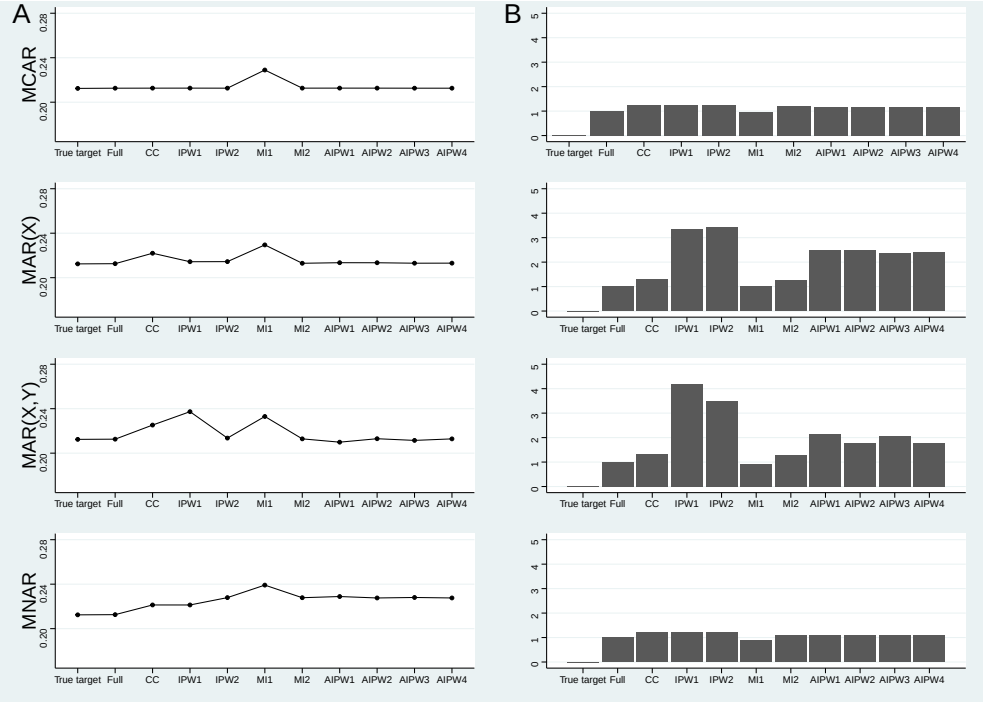

Figure 2.4: Simulation results of mean and relative SD of BS for existing model M_2: best model based on just X_1, X_2. Column A denotes mean BS. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

generally lead to an inferior model. As shown in Fig.2.3 and Fig.2.4 the full model results do achieve the target true value for M_2, and they represent the best that could be achieved for M_2. The relative performance of the various MI, IPW and AIPW methods for the handling the missing data compared to the full model results are quite similar to those shown in Fig.2.1 and Fig.2.2, both for bias and SD.

We also considered using a poor external model M_3 with wrong coefficients. The results are shown in Fig.2.5 and Fig.2.6. Again in comparison with full data analysis, the MI2, IPW2, AIPW2 and AIPW4 appear to give no bias, except in the MNAR case. The variability of the MI2 method is the smallest.

For the scenario when external data and internal data have different prevalence, we consider an existing model with the intercept=1 while the other coefficients are the same as the true model. The changed intercept in M_4 has no influence on the

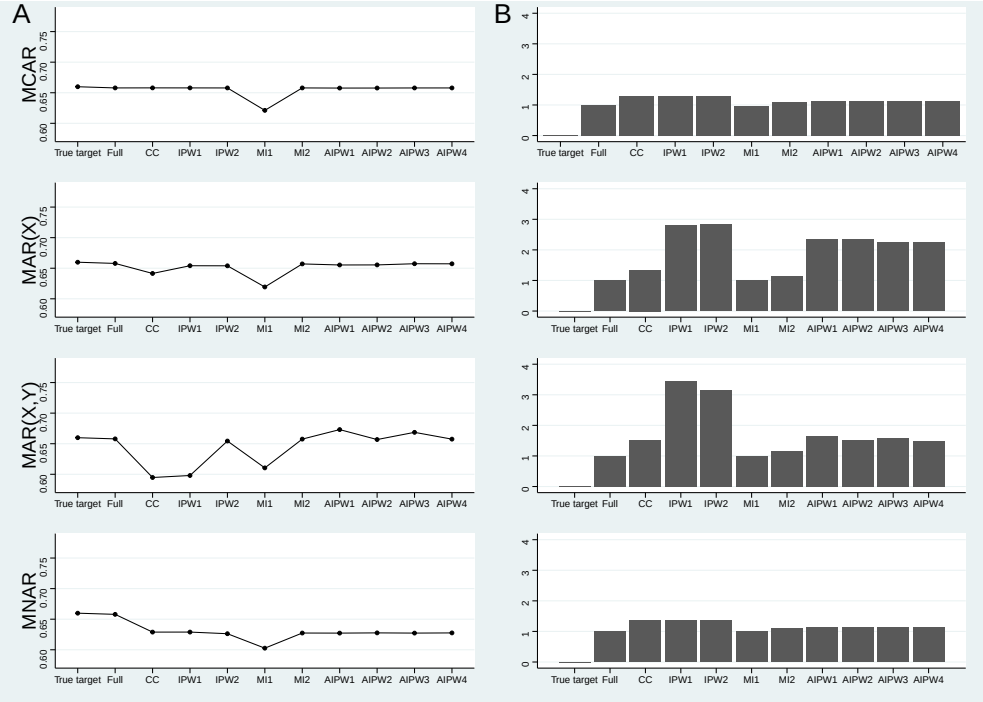

Figure 2.5: Simulation results of mean and relative SD of AUC for existing model M_3: poor model based on X_1, X_2, X_3. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

AUC compared to the true value, since changing the intercept does not change the discrimination ability. The results are identical to those shown in Fig.2.1. The values of BS increased compared to situation M_1. As shown in Fig.2.7 the relative merits of the MI, IPW and AIPW methods are similar to the other scenarios.

Overall, for the situations considered in this study, considering both bias and variability the best methods are MI2 and AIPW4. For correlated covariates, the conclusions are the same (see Appendix). With multiple missing covariates, the findings are broadly similar, but with some differences depending on the missingness pattern. The simulation results shown in the Appendix, suggest that here MI2 is the best method.

We note that the model used to impute the missing X in MI2 and create X^* in AIPW3 and AIPW4 is slightly misspecified. Although it does regress X_1 on

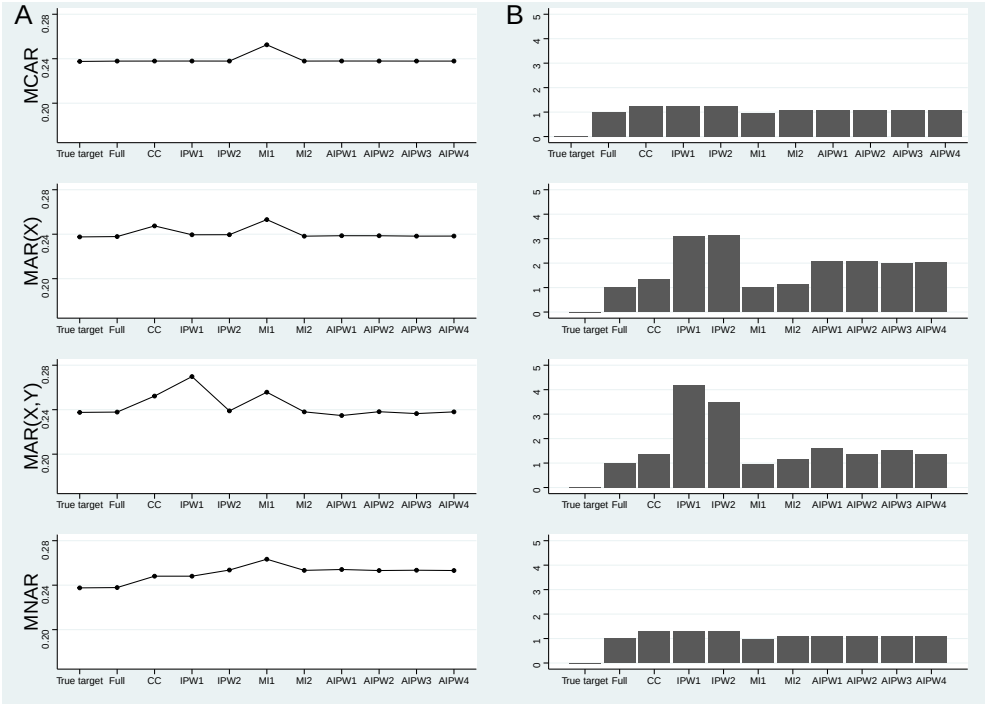

Figure 2.6: Simulation results of mean and relative SD of BS for existing model M_3: poor model based on X_1, X_2, X_3. Column A denotes mean BS. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

X_2, X_3 and Y, the assumed linear model is not the same as the true distribution for $X_1|X_2, X_3, Y$ based on how the data was generated from the true model. Furthermore, we use $I(\beta X_i^* > \beta X_j^*)$ to approximate $E[I(\beta X_i > \beta X_j)]$ as in the proof which is an approximation to a doubly robust method. This, together with the fact that the AIPW method is an approximation to the doubly robust estimator, may explain the small bias in the AIPW3 method for the MAR(X_2, Y) case, because in fact neither the weight model nor the imputing model is correctly specified. However the misspecified imputing model does not give any noticeable bias for the MI2 method.

2.4 Application

In this section, we applied the proposed methods to evaluate the performance of an existing model for the risk of recurrence in men with Prostate Cancer. The Cancer

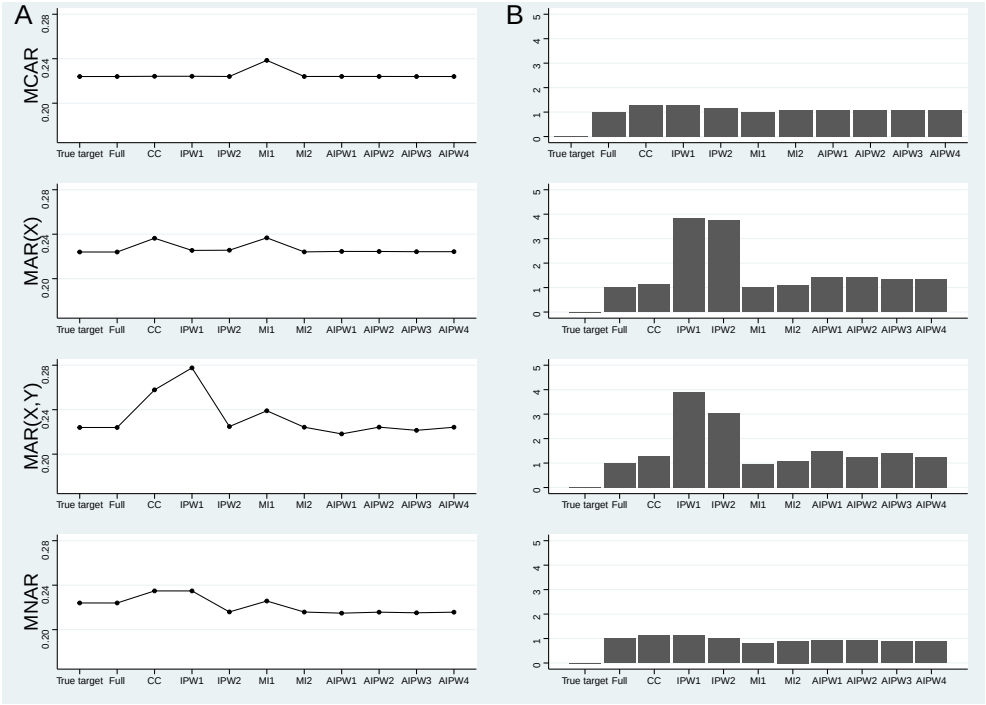

Figure 2.7: Simulation results of mean and relative SD of BS for existing model M_4: different intercept model. Column A denotes mean BS. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

of the Prostate Risk Assessment (CAPRA) score was published in 2005 and was based on an initial cohort consisting of >1400 men from the University of California, San Francisco (UCSF) [9]. A Cox proportional hazards regression model identified age, pretreatment PSA, Gleason score, percentage of biopsy cores positive for cancer (PPC), and clinical stage as significant factors associated with biochemical recurrence (BCR) or secondary treatment. Based on the results of the Cox analysis, points were assigned as in Table 2.2 to indicate relative risk. For each patient the points would be added to give an overall CAPRA score. The CAPRA score ranges from 0 to 10, and every 2-point increase in the score represents an approximate doubling of the risk. The distribution of the score and the 3 year recurrence-free survival (RFS) rate were reported in the publication, and are shown in Table 2.3. The AUC can be calculated from the CAPRA score itself, but the BS requires the predicted probabilities from

Table 2.3.

Table 2.2: CAPRA score

Variable	Level	Points
PSA	2.0-6	0
	6.1-10	1
	10.1-20	2
	20.1-30	3
	>30	4
Gleason Score (Primary/Secondary)	1-3/1-3	0
	1-3/4-5	1
	4-5/1-5	3
T stage	T1/T2	0
	T3a	1
Percent positive biopsy	<34%	0
	≥ 34%	1
Age	<50	0
	≥50	1

Table 2.3: CAPRA score distribution and predicted probabilities derived from the CAPRA score.

CAPRA Score	CAPRA score distribution	3-Yr RFS rate
0-1	27.9%	0.91
2	30.0%	0.89
3	20.6%	0.81
4	10.8%	0.81
5	5.8%	0.69
6	3.0%	0.54
7 or Greater	2.0%	0.24

We sought to estimate the performance of CAPRA using a separate dataset from the Mayo Clinic. The 1268 patients were treated with surgery between 2008 and 2012 and all patients before 2010 and half patients later were missing PPC values. So in total 90% of the patients were missing PPC. We considered 3-year RFS as a binary outcome. We included in our analysis all men who were followed more than 3 years or developed progression in 3 years. In total, 314 of the 1268 patients had a recurrence

in 3 years. To validate the prediction of CAPRA score, we compared the CAPRA score with the outcome to get the AUC, and compared the observed RFS rate for each CAPRA score as in Table 2.3 with the outcome to get Brier score. Because 90% patients have missingness in PPC, we used PSA, Gleason Score, T-stage, Age and/or the outcome to build the weight model for missingness and the imputation model of PPC in the IPW, AIPW, and MI methods. In the data analysis, `mice()` in `R` with logistic regression is used to implement MI for the missing binary PPC. `glm()` with logistic link was used to build weight models and `glm()` with logistic link was used to calculate the predicted PPC in AIPW. A bootstrap was used to give 95% confidence intervals for AUC and BS.

Fig.2.8 shows the analysis results of different methods. The AUC ranged from 0.73 to 0.79, which is similar to other external validation studies of the CAPRA score for which the c-index for BCR ranged from 0.66 to 0.81 [2]. On the other hand, the BS values were around 0.16 except for complete case analysis and IPW with the weight model excluding the outcome variable, which were above 0.4. The complete case analysis and IPW methods have much wider confidence intervals, while the MI and AIPW methods have comparable confidence intervals. Little's test was used and indicated the missingness is not MCAR ($p<0.001$) [26], thus complete case analysis is not an optimal choice. The Odds Ratio of PPC not missing and RFS observed was 24.1, indicating the missingness was strongly related to the outcome. Thus the methods in which the weight model includes the outcome should be more reliable. The imputing model of PPC was built only on the 10% of patients with non-missing data and was used to impute the other 90% later on, and there could be a large variation in the model, which could explain the ignorable difference between the two MI methods with or without outcome. The results for AUC and BS are different, probably because some CAPRA scores have the same RFS rate.

These results indicate the approaches to handle missing data can result in fairly

large variation in model performance estimates. Based on the theoretical considerations and the simulation results, we believe the results from MI2 and AIPW4 are the best to use, and they give very similar estimates for both BS and AUC in this example.

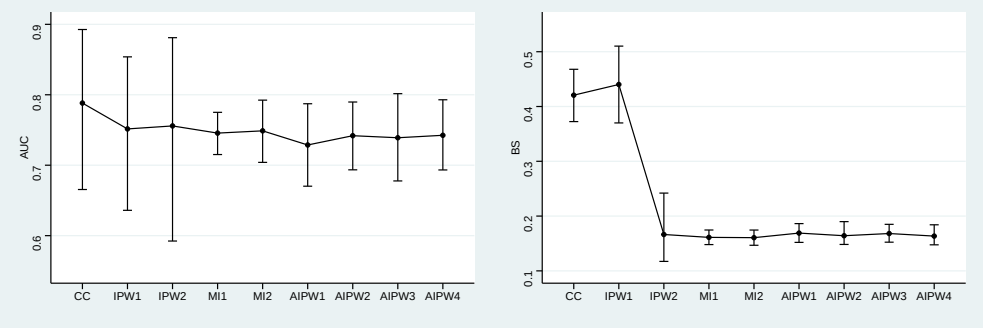

Figure 2.8: Varying estimates of mean and 95% confidence interval of AUC and Brier Scores for prostate cancer example, based on how missing data are handled

2.5 Discussion

We developed new AIPW estimators for predictive model performance metrics in the setting of missing data. This AIPW approach is shown to have good properties. We note that an AIPW estimator of the AUC has been previously proposed [29], but for a different setting with auxiliary variables. Adapting this published approach to our setting does not lead to equation (2.13), but rather an estimator with weights in the denominator as in equation (2.10). When the weight model is correctly specified and with assumed independence of cases and controls, the expectation of the denominator in equation (2.10) is equivalent to the denominator in equation (2.13) .

When there are missing observations in the internal data, MI and IPW can both be used to obtain unbiased estimates of BS and AUC if the imputation model or weight model is correctly specified. When the missingness doesn't depend on Y, IPW doesn't need to include Y in the weight model, while MI does need to include Y in the

imputation model. When the missingness depends on Y, both IPW and MI need to include Y. The outcome variable should be included in the imputation model under all scenarios, because it provides information of the missing covariates. For IPW, the outcome only need to be included in weight model if the missingness depends on outcome in order to get the correctly specified weight model. The findings in this paper clearly support inclusion of the outcome variable Y in models that handle the missing covariates when evaluating an existing prediction model. Thus overall, even though in some situations for the IPW and AIPW methods it is not necessary, very little harm arose from including Y and there is the potential for considerable gain.

Our simulation results suggest that AIPW can be more efficient than IPW, and also obtain approximate double robustness to mis-specification of the weight model or the imputing model. Even when both models are mis-specified, resulting estimates are still less biased than IPW or MI with the wrong weight model or imputing model. Under all scenarios, MI has the best efficiency comparing to full data analysis. Under MCAR, AIPW has the same efficiency as MI, while under MAR, AIPW is less efficient than MI.

One limitation of the IPW and AIPW methods is when there are multiple covariates missing. In this situation there are different possible ways in which the weight model and the imputation model can be constructed. In he specil cases of blocked missingness or monotone missingness there are natural ways to construct these models, and in the simulation study we found similar performance to that of the situation with a single missing covariate. When the missingness is scattered there are more choice of how to implement the imputation model, and our simulation results suggest that AIPW can in fact be a less desirable method than IPW. With multiple missing covariates the MI methods are still relatively easy to apply by using the chained equation approach to impute the missing values sequentially, and the simulation results suggest it is clearly more efficient.

The problem we consider in this paper is how to estimate the correct AUC and BS for a different population than the one that was used to develop the prediction model, when (i) we do not have access to the data that was used to develop the model and (ii) the dataset we have from the different population has some missing covariate values. There are a broad set of problems associated with missing covariates and risk prediction models. One is how to develop a model, for which a much cited reference is Moons et al [33]. Another set of problems is how to implement an existing risk prediction model for an individual subject when that subject has some missing covariates, and also will not have the outcome known. Different situations and possibilities exist here. The model developer may have set up methods to use in the case of missing data, such as 2^k different models, one for each pattern of missingness. The user of the model may simply try a range of values for the missing variables, to give a range of predicted probabilities, analogous to sensitivity analysis. If the user of the model has access to the training data, then the question becomes how to make use of these data. Alternatively, the user of the model may have access to their own dataset, with information on both the covariates and outcomes for people in this dataset, and the individual subject can be considered as coming from the same population as this dataset, then the question again is how to make use of these data. These challenges have received limited attention in the statistical literature [32, 20], but have been expounded upon in a recent publication [18].

Another situation worthy of study, is how to evaluate an existing prediction model, in a different population, when that different population does not have measured one of the needed input variables for the prediction model. This would seem to be an impossible task, unless extra information is available, either in the form of additional data or knowledge of the joint distribution of the missing variable with the other variables.

CHAPTER III

A Utility Approach to Individualized Optimal Dose Selection Using Biomarkers

3.1 Introduction

The goal of personalized medicine is to give the right treatment to the right patient at the right dose using all we know about the patient. Available knowledge about individual patients is increasingly including biomarkers which allow personalizing treatment decisions. One approach to personalized medicine is to identify the right patient for a given treatment. For example, in the setting of a single binary outcome and two potential treatments, Foster et. al [12] proposed a "Virtual Twins" method involving predicting response probabilities for treatment and control "twins" for each subject by random forest, and then using regression or classification trees to identify subgroups of patients with large positive treatment effect estimates. A related but different approach is to identify the right treatment for a patient, often referred to as optimal treatment regimes (OTR). A treatment regime is defined as the function that maps a patient's covariate vector to one of the treatment choices. One approach to identify an OTR is a 2-step method that involves building a model for conditional expectation of the outcome given treatment as the first step, then maximizing the mean expected reward to get the optimal treatment for each subject. In an alternative

approach, rather than modeling the marginal outcome, outcome weighted learning (OWL) methods maximize the reward from following a treatment regime directly, which is equal to the expected outcome in a subset of patients who actually followed that regime, inversely weighted by the probability of being assigned to the regime [55]. Maximizing the reward with respect to the treatment regime is equivalent to minimizing the expectation for patients who did not follow the regime, and can be interpreted as minimizing the weighted classification error in a classification problem [53]. Zhang et. al [54] also proposed the doubly robust augmented inverse probability weighted estimator (AIPWE) in which an outcome model is combined with a model for the probability of a treatment which is important when analyzing observational data. The OWL method has been extended to continuous treatment dose settings such as optimal dose finding [5].

In many settings, it is not possible to describe a patient's outcome using a single variable. For example, in oncology, it is typical to describe patient outcomes in terms of toxicity and efficacy variables. Several strategies have been proposed for identifying an optimal treatment or dose based on the trade-off between efficacy and toxicity. Thall and Cook [46] proposed using efficacy-toxicity trade-off contours that partition the two-dimensional outcome probability domain such that efficacy-toxicity pairs on the same contour are equally desirable. Dose could then be selected to maximize desirability. More commonly, a utility matrix is elicited from clinicians by assigning numerical utilities to each possible bivariate outcome. The optimal dose is then defined as the value maximizing the posterior mean utility [15].

Guo and Yuan [15] proposed a Phase I/II trial design incorporating biomarkers in which the optimal dose for an individual patient is selected to maximize utility. A joint model of ordinal toxicity and efficacy outcomes is specified and canonical partial least squares are used to extract a small number of components from the covariate matrix containing dose, biomarkers, and dose-by-biomarker interactions. Wang et. al

[49] proposed two approaches to identify a personalized optimal treatment strategy that maximizes clinical benefit under a constraint on the average risk in the situation of a binary treatment option and continuous outcomes.

In this paper we propose a utility based method to estimate optimal doses for individual patients in the setting of binary efficacy and toxicity outcomes. To allow for potentially large numbers of biomarkers and patient factors we utilize l_1-penalty via LASSO [47]. At the individual level, we find the optimal dose by maximizing utility functions defined as the probability of efficacy minus the weighted probability of toxicity, which is equivalent to a utility matrix [40]. The weight term in the utility equation could be elicited from clinicians to quantify the relative undesirability of toxicity relative to lack of efficacy. Alternatively it can be viewed as a tuning parameter selected to achieve a desired overall (at the population level) rate of toxicity. In the vast majority of oncology treatments and many other disease settings, both efficacy and toxicity outcomes are monotonically linked to increasing dose. While "flat" curves are common [38], it is uncommon for increasing dose to lead to decreased toxicity or efficacy. We note that monotonicity may not hold for outcomes such as progression free survival which include death as an event, since they are potentially a consequence of either toxicity or lack of efficacy. When estimating outcomes as a function of dose only it is often not necessary to impose this constraint. However, when including many potential dose*biomarker interactions, it is likely that some patients will be estimated to have decreasing toxicity or efficacy with increasing dose due to statistical noise. To prevent this and to improve efficiency we propose a method that constrains the estimated dose-efficacy and dose-toxicity relationships to be non-decreasing for all patients. We call this constrained LASSO, which can be solved by decomposition and quadratic programming [16] and alternating direction method of multipliers (ADMM) [13]. In section 3, we report results of a simulation study and in section 4 we illustrate the proposed methods using a dataset of patients with lung

cancer treated with radiation therapy.

3.2 Method

3.2.1 Binary Outcome Setting

We assume the available data, $(x_i, d_i, E_i, T_i), i = 1, ..., n$, comprises n independent and identically distributed copies of (x, d, E, T), where x, a p-dimensional centered vector of subject-specific features, $d \in [-1, 1]$ denotes continuous dose of treatment, E is the binary efficacy outcome, and T is the binary toxicity outcome. A large probability of E and small probability of T is preferable.

An individualized dose rule is the map from x to the dose domain: $\mathcal{F} : \mathbb{R}^p \to [-1, 1]$. Under $\mathcal{F}$, a patient with covariate x is recommended to dose $d = \mathcal{F}(x)$. For any treatment rule $\mathcal{F}$, the population expected efficacy and toxicity are $\mathbb{E}^{\mathcal{F}}(E)$ and $\mathbb{E}^{\mathcal{F}}(T)$. Our goal is to estimate an individualized dose rule that maximizes the population expected efficacy while controlling the overall expected toxicity under some tolerance level, that is

$$\max_{\mathcal{F}} \quad \mathbb{E}^{\mathcal{F}}(E) \text{ , subject to } \mathbb{E}^{\mathcal{F}}(T) \leq \tau, \tag{3.1}$$

where τ is the pre-specified maximal tolerance level of average toxicity.

Define $\delta_E(x_i, d_i)$ =P$(E = 1|d_i, x_i)$−P$(E = 1|d_i = -1, x_i)$, $\delta_T(x_i, d_i)$ =P$(T = 1|d_i, x_i)$−P$(T = 1|d_i = -1, x_i)$. $\delta_E(x_i, d_i)$ and $\delta_T(x_i, d_i)$ can be interpreted as the difference in expected efficacy and toxicity outcomes for a patient if treated at the lowest dose $(d_i = -1)$ or some higher dose (d_i). Let $\mathbb{E}_x$ to denote the population average of the function across the distribution of x. After introducing the Lagrange multiplier, solving equation (3.1) is equivalent to

$$\max_{\mathcal{F}} \quad \mathbb{E}_x[\delta_E\{x_i, \mathcal{F}(x_i)\} - \theta\delta_T\{x_i, \mathcal{F}(x_i)\}], \tag{3.2}$$

where $\theta > 0$ is chosen such that $\mathbb{E}_x[\delta_T\{x_i, \mathcal{F}(x_i)\}] \leq \tau - \mathbb{E}_x[\mathrm{P}(T = 1|d_i = -1, x_i)]$. The expression in equation (3.2) is a utility function quantifying the trade-off between efficacy and toxicity. By fitting separate models for E and T using methods such as logistic regression via maximum likelihood or constrained LASSO as described below, we can calculate the utility values for individual patients over the range of possible dose values, and calculate the dose rule that maximizes equation (3.2).

Consider the model $\mathrm{logit}\{\mathrm{P}(Y = 1)\} = f(x, d, \beta) = \beta_0 + W\beta$ between outcome Y, i.e. E or T, and covariates including biomarkers x, dose d, and dose-biomarker interactions dx, i.e. $W = (x, d, dx)$. To fit the model we use the generalized LASSO with l_1-penalty on the log-likelihood and no penalty on β_0. To enforce a non-decreasing relationship of efficacy and toxicity with dose, we add constraints on derivatives with respect to d to be non-negative, i.e., $\frac{\partial}{\partial d} f(x_i, d, \beta) \geq 0$ for all x_i. We call this method constrained LASSO (cLASSO), for which the constraint can be written as $C\beta \geq 0$, where C is a $n \times (2p + 1)$ matrix of $[\mathbf{0}_{n*p}, 1_{n*1}, \boldsymbol{x}_{n*p}]$. Then the cLASSO method is to

$$\text{minimize} \quad -\log L(\beta, W, Y) + \lambda\|\beta\|_1 = -\sum_{i=1}^{n} [Y_i(\beta_0 + W_i\beta) - \log\{1 + \exp(\beta_0 + W_i\beta)\}] + \lambda\|\beta\|_1 \tag{3.3}$$

subject to $C\beta \geq 0$.

To solve (3.3) we decompose β into its positive and negative part, $\beta = \beta^+ - \beta^-$, as the relation $|\beta| = \beta^+ + \beta^-$ handles the l_1 penalty term. Let $W^* = (W, -W)$, and $\beta^* = (\beta^{+T}, -\beta^{-T})^T$. By plugging these into (3.3) and adding the additional non-negativity constraints on β^+ and β^-, the constrained LASSO is formulated and can be solved, for example, by `spg()` in `R`, which uses the spectral projected gradient method for large-scale optimization with simple constraints. That is,

$$\text{minimize} \quad -\sum_{i=1}^{n} [Y_i(\beta_0 + W_i^*\beta^*) - \log\{1 + \exp(\beta_0 + W_i^*\beta^*)\}] + \lambda 1_{4p+2}{}^T \beta^* \tag{3.4}$$

subject to $(C, -C)\beta^* \geq 0, \beta^+ \geq 0, \beta^- \geq 0$.

The derivative of (3.4) with respect to (β_0, β^*) is

$$-\sum_{i=1}^{n} \left\{ Y_i(1, W_i^*) - \frac{\exp(\beta_0 + W_i^*\beta^*)}{1 + \exp(\beta_0 + W_i^*\beta^*)}(1, W_i^*) \right\} + \lambda(0, 1_{4p+2}{}^T) \tag{3.5}$$

The minimizer to (3.4) always satisfies $\beta_j^+ \beta_j^- = 0$ for $j = 1, ..., 2p + 1$, as shown in the Appendix. We use 10-fold cross-validation(CV) to choose λ to minimize the CV deviance.

For any fixed value of theta, and using the above estimated models of efficacy and toxicity, we can find the optimal dose for each patient that maximize (3.2), then we use grid search to find the smallest θ achieving the constraint on toxicity. Specifically the algorithm is as follows:

1. Set a grid $0 = \theta_1 < \theta_2 < ... < \theta_K$

2. For each $m = 1, ..., K$:

(a) set $\theta = \theta_m$

(b) For each subject $i = 1, ..., n$ with covariate x_i:

calculate $d_i^{opt} = \text{argmax}(\delta_E(x_i, d) - \theta\delta_T(x_i, d))$, estimate $\text{P}(E = 1|d_i^{opt}, x_i)$ and $\text{P}(T = 1|d_i^{opt}, x_i)$

3. Select the smallest $\hat{\theta}$ such that $\mathbb{E}_x[\text{P}(T = 1|d_i^{opt}(\theta), x_i)] \leq \tau$. Then $d_i^{opt}(\hat{\theta})$ is the estimated optimal dose for patient i.

For binary outcomes under the logistic link function, both $\delta_E(x, d)$ and $\delta_T(x, d)$ are functions involving the intercept, main effect of x, as well as dose related covariates d and dx. Estimation of d^{opt} at given θ for each subject is solved by one-dimensional optimization using `optimize()` in `R`, which uses a combination of golden section search and successive parabolic interpolation. Because of the non-decreasing dose-efficacy and dose-toxicity relationship, a larger θ will recommend a smaller d^{opt}, and

the corresponding population average efficacy and toxicity will be smaller. So the smallest θ achieving the constraint on average toxicity will achieve the largest average efficacy. The range and size of the grid can be pre-specified and should include a range of feasible values. In our simulation and data example, we used a range of 0.01 to 4, in steps of 0.001. We note that for the determination of d_i^{opt}, we consider the subject level $E-T$ trade-off, while for the determination of θ, we look at the population level $E-T$ trade-off.

3.2.2 Multiple Outcome Setting

In some applications there are multiple toxicity outcomes which must be considered and balanced against efficacy when selecting treatment dose. Without loss of generality, we consider two different toxicity outcomes T_1, T_2 and the goal is to

$$\max_{\mathcal{F}} \quad \mathbb{E}^{\mathcal{F}}(E) \text{ , subject to } \mathbb{E}^{\mathcal{F}}(T_1) \leq \tau_1, \mathbb{E}^{\mathcal{F}}(T_2) \leq \tau_2 \tag{3.6}$$

where τ_1, τ_2 are the pre-specified maximal tolerance levels of average toxicity for each toxicity outcome.

Define $\delta_E(x_i, d_i)$ =P$(E = 1|d_i, x_i)$−P$(E = 1|d_i = -1, x)$, $\delta_{T_1}(x_i, d_i)$ =P$(T_1 = 1|d_i, x_i)$−P$(T_1 = 1|d_i = -1, x_i)$, $\delta_{T_2}(x_i, d_i)$ =P$(T_2 = 1|d_i, x_i)$−P$(T_2 = 1|d_i = -1, x_i)$. Then equation (3.6) is equivalent to

$$\max_{\mathcal{F}} \quad \mathbb{E}_x[\delta_E\{x_i, \mathcal{F}(x_i)\} - \theta_1\delta_{T_1}\{x_i, \mathcal{F}(x_i)\} - \theta_2\delta_{T_2}\{x_i, \mathcal{F}(x_i)\}], \tag{3.7}$$

where $\theta_1 > 0, \theta_2 > 0$ are chosen such that $\mathbb{E}_x[\delta_{T_1}\{x_i, \mathcal{F}(x_i)\}] \leq \tau_1 - \mathbb{E}_x[\text{P}(T_1 = 1|d_i = -1, x_i)]$, and $\mathbb{E}_x[\delta_{T_2}\{x_i, \mathcal{F}(x_i)\}] \leq \tau_2 - \mathbb{E}_x[\text{P}(T_2 = 1|d_i = -1, x_i)]$.

We specify parametric logistic models for E, T_1, T_2 as functions of biomarkers, dose, and dose-biomarker interactions. Denote the parameter estimates from those

logistic models as $\hat{\beta}_E, \hat{\beta}_{T_1}, \hat{\beta}_{T_2}$. We propose a random walk and Metropolis algorithm to select θ_1, θ_2 to achieve the constraints on toxicity. The algorithm is as follows:

1. Set a chain length, B, fix $\sigma^2 > 0$ and initialize $\theta^0 = (\theta_1^0, \theta_2^0)$ to a starting value that makes $\mathbb{E}_x[\mathrm{P}(T_1 = 1|d_i^{opt}(\theta^0), x_i)] \leq \tau_1, \mathbb{E}_x[\mathrm{P}(T_2 = 1|d_i^{opt}(\theta^0), x_i)] \leq \tau_2$.

2. For $b = 0, ..., B$:

(a) Generate $\tilde{\theta}^{b+1} \sim N(\theta^b, \sigma^2 \boldsymbol{I})$ and $\tilde{\theta}^{b+1} > \mathbf{0}$

(b) For each subject $i = 1, ..., n$ with covariate x_i:

compute $d_i^{opt} = \mathrm{argmax}\{\delta_E(x_i, d) - \tilde{\theta}_1^{b+1}\delta_{T_1}(x_i, d) - \tilde{\theta}_2^{b+1}\delta_{T_2}(x_i, d)\}$, estimate $\mathrm{P}(E = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)$, $\mathrm{P}(T_1 = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)$, and $\mathrm{P}(T_2 = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)$

(c) Compute $q = \min[1, \mathbb{E}_x\{\mathrm{P}(E = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)\}/\mathbb{E}_x\{Pr(E = 1|d_i^{opt}(\theta^b), x_i)\}]$

(d) Generate $U \sim U(0, 1)$;

if $\mathbb{E}_x[\mathrm{P}(T_1 = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)] \leq \tau_1, \mathbb{E}_x[\mathrm{P}(T_2 = 1|d_i^{opt}(\tilde{\theta}^{b+1}), x_i)] \leq \tau_2$, and $U \leq q$, set $\theta^{b+1} = \tilde{\theta}^{b+1}$; otherwise, set $\theta^{b+1} = \theta^b$

3. After generating a chain $(\theta^0, ..., \theta^B)$, we select the θ^k that leads to the largest value of $\mathbb{E}_x[\mathrm{P}(E = 1|d_i^{opt}(\theta^k), x_i)]$ as the optimal solution, and the $d_i^{opt}(\theta^k)$ is the optimal dose for patient i.

In stage 2(b) in the above algorithm, d^{opt} at given θ for each subject is solved by one-dimensional optimization using `optimize()` in `R`. The variance of the proposal distribution σ^2 in stage 2(a) is chosen to make the acceptance proportion between 0.25 and 0.5. When there are multiple constraints we found that the random walk and Metropolis algorithm is more efficient than using a finite grid search over the multiple dimensions of θ. In our experience, as long as the chain is long enough, the maxima of the population average efficacy will be achieved. This can be checked by running the algorithm in parallel for different initial choices of θ^0. It is noted that there is no guarantee that both toxicity constraints will be met at the boundary.

3.3 Simulation Studies

In this section, we performed numerical studies to investigate the performance of the proposed method under different settings. We simulated five i.i.d covariates, $x_1, ..., x_5$ from a standard normal distribution, d from Uniform(-1,1), and then generated N=200 binary outcomes E and T from the regression models

$$\text{logit}(\text{P}(E = 1)) = \beta_{0,E} + W_E\beta_E,$$
$$\text{logit}(\text{P}(T = 1)) = \beta_{0,T} + W_T\beta_T,$$

where $W_E = W_T = (x, d, dx)$. A range of scenarios for β were considered, but we first describe scenario 0, as given in Table 4.1. For scenario 0 $(\beta_{0,E}, \beta_E)$ =(0, 1, 0, 0, 0, 0, 1, .4, .4, .4, -.8, 0), and $(\beta_{0,T}, \beta_T)$ =(-1.386, -1, 0, 0, 0, 0, 1, -.4, -.4, -.4, .8, 0). In generating x's, we also applied the constraints that x must satisfy $1 + 0.4x_1 + 0.4x_2 + 0.4x_3 - 0.8x_4 > 0$ and $1 - 0.4x_1 - 0.4x_2 - 0.4x_3 + 0.8x_4 > 0$ to reflect the non-decreasing dose-efficacy and dose-toxicity curves for all subjects. This excludes up to 35% of the originally simulated observations.

To illustrate the utility approach to dose selection, we plotted individual level E-T trade-off for three different subjects in Fig.3.1 and population level E-T trade-off in Fig.3.2. Different dose-efficacy and dose-toxicity curves among subjects result in selection of different optimal dose values across θ.

For variable selection, we forced the main effect for dose to be selected by removing its associated parameter from the penalty term and only consider the selection of covariates and dose*covariate interactions. The methods we compared are forward selection (FS), regular LASSO, cLASSO and fixed dosing (FD) in which dose only logistic models were fit. FS was implemented by `step()` in `R` using AIC as criteria. Regular LASSO was implemented by `glmnet()` in `R` with 10-fold CV.

The boxplots in Fig.3.3 shows the average efficacy from the above methods with the same toxicity constraint, from which we see that cLASSO has higher average

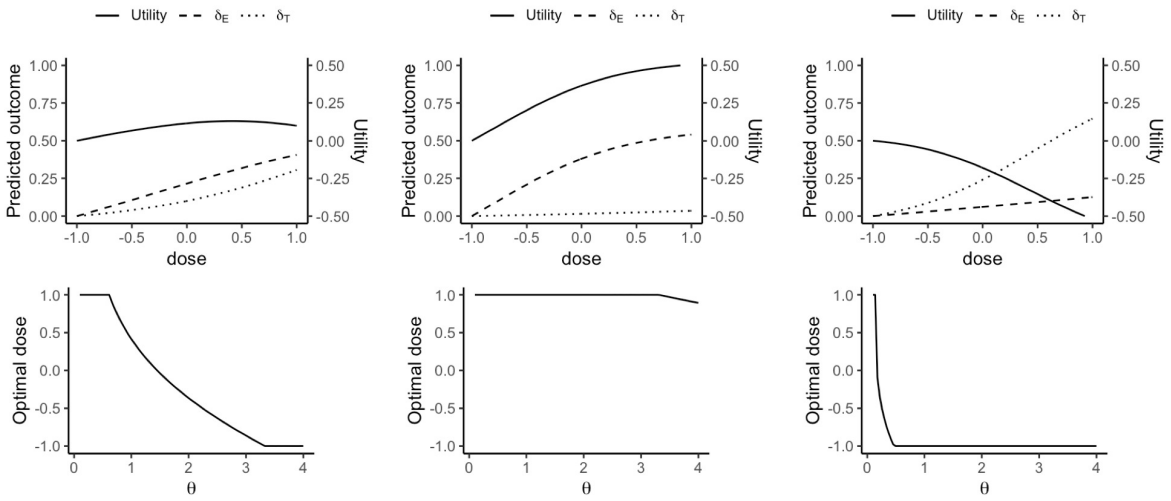

Figure 3.1: Top: Individual level E-T plot with choice of dose with theoretical β_E, β_T for three subjects. The utility curve uses $\theta = 1$. Bottom: Individual level optimal dose as a function of θ for the same three subjects.

efficacy compared to the other methods, especially fixed-dose. We also calculate the theoretical improvement from using the true models which would only be known in a simulation study setting. The improvement, defined as proportion of possible gain compared to the gain from FD to theory, $\{\mathbb{E}^{\mathcal{F}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}/\{\mathbb{E}^{\mathcal{F}_{Theory}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}$, for FS, regular LASSO, and cLASSO is 0.513, 0.518, 0.589, respectively. cLASSO has higher efficacy than LASSO for 69.0% of the simulated datasets.

We also considered a null case in which there are no covariates or dose*covariates interactions with result shown in Fig.3.4, so that the dose effects are the same across all subjects. The average efficacy with toxicity constrained at 0.2 for theory, FS, regular LASSO, cLASSO and FD are 0.503, 0.470, 0.490, 0.491, 0.499, respectively. With no effect of covariates, the dose-only model (FD) is as good as the theory, and the models with covariates included have slightly worse performance than the dose-only model. Among the modeling approaches, cLASSO has better performance than

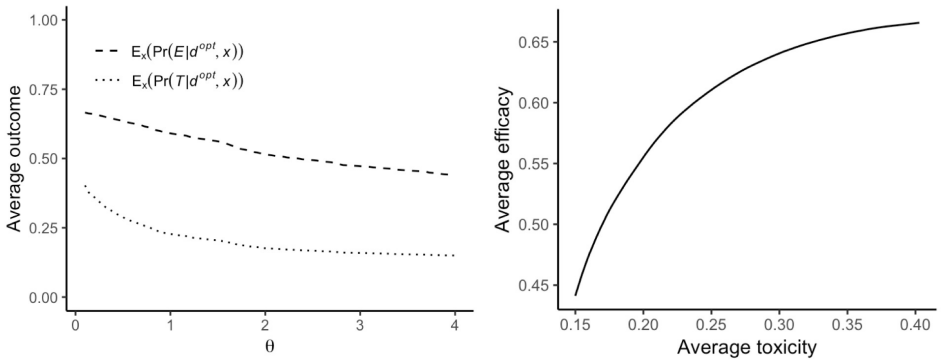

Figure 3.2: Left: Population level E-T plot with choice of θ with theoretical β_E, β_T, Right: Population level E-T trade-off at different toxicity tolerance levels.

FS.

A few other scenarios were considered: In scenario 1 there are only main effects of covariate and no dose-covariate interactions, and the main effects of the same x are in opposite directions in efficacy and toxicity models. In scenario 2, the dose-covariate interaction effects are the same in efficacy and toxicity models, but the main effects of covariates are different. In scenario 3 and 4, there are 15 and 45 additional noise covariates added to increase p from 5 to 20 and 50, to examine the performance with high dimensional data. In scenario 5 and 6, with the same coefficients as in scenario 0 and 3, the sample size increased to 400. In practice, the covariates may be highly correlated resulting in multicollinearity. In scenario 7, x's are not independent and correlation among x_1, x_2, x_3 are 0.6. We also considered several situations in which the logistic regression model with linear effects is mis-specified. In scenario 8, the true effect of covariate x_4 is stepwise at 0, i.e., the effect only exits for $x_4 > 0$. But when fitting models, it is mis-specified as linear. In scenario 9, the true models have $exp(x_4)$ as the covariate, but in the fitted models x_4 is used, which is mis-specified. In scenario 10, an interaction of x_2 and x_3 is included as main effect in both true models

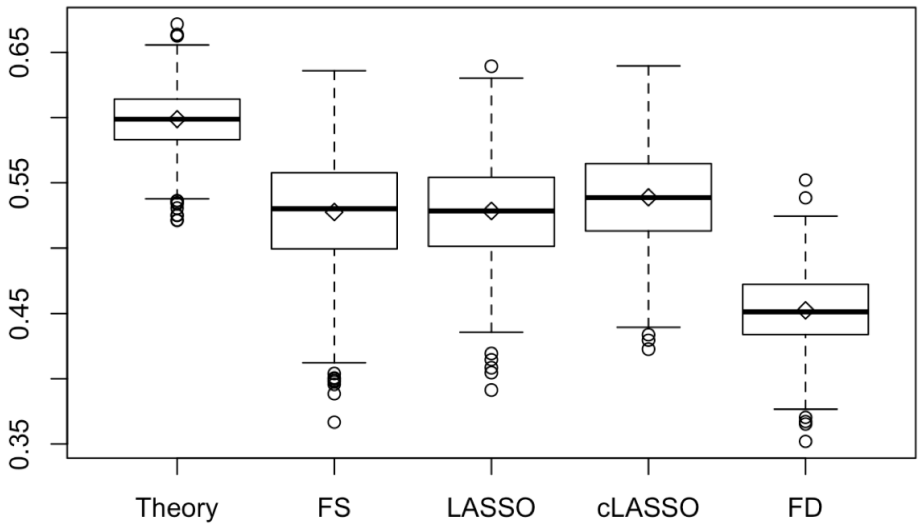

Figure 3.3: Simulation results for scenario 0. Boxplot of average efficacy with same toxicity for 1000 simulation trials. The compared methods are theory with true coefficients; FS: Forward Selection; LASSO; cLASSO: constrained LASSO; FD: Fixed Dosing. All methods are constrainted at P(T)=0.20. Means are 0.599, 0.528, 0.528, 0.539, 0.452, respectively.

for efficacy and toxicity, but in the fitted models this interaction is not included.

Table 3.1 showed the simulation results with the above setting. We also considered another setting with more covariate main effects and fewer dose-covariate interactions, and with the non-decreasing constraints, 12% of the simulated observations were excluded. In scenario 1 and 2, when the dose related coefficients for efficacy and toxicity models are the same or 0, the main effects of covariates still played a role in the optimal dose finding with the logistic link, and cLASSO still has better performance than the other methods. In scenario 3 and 4, with the increased number of noise covariates, the magnitude of improvement decreased, but the cLASSO still performs better than the other methods. In scenario 5 and 6, with the larger sample sizes, the magnitude of improvement increased, and cLASSO outperforms the other methods. In scenario 7, with correlated covariates where the performance of LASSO is known to be suboptimal, cLASSO still performs better than LASSO and FS. In scenario

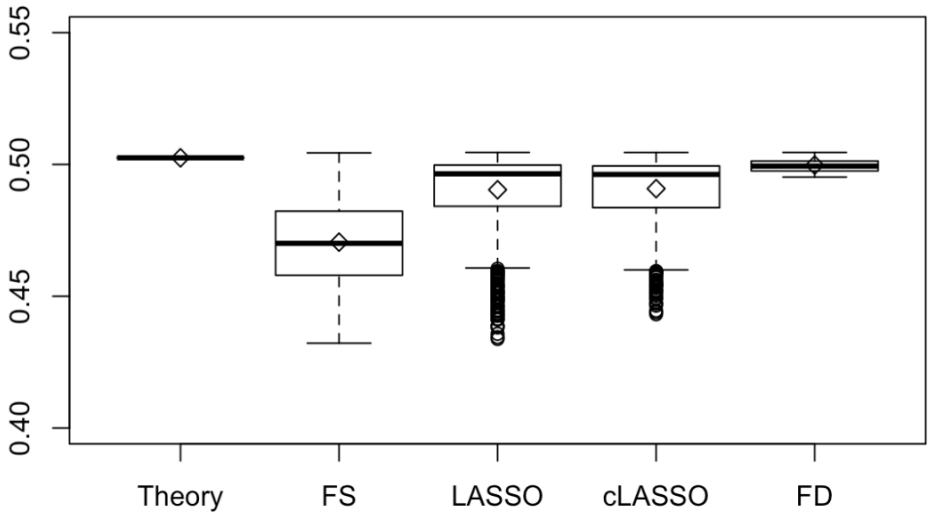

Figure 3.4: Simulation results for the null scenario. Boxplot of average efficacy with same toxicity for 1000 simulation trials. The compared methods are theory with true coefficients; FS: Forward Selection; LASSO; cLASSO: constrained LASSO; FD: Fixed Dosing. All methods are constrainted at P(T)=0.20. Means are 0.503, 0.470, 0.490, 0.491, 0.499,, respectively.

8, 9, and 10, when the logistic regression model with linear effects is mis-specified, all the methods have smaller magnitude of improvement, but cLASSO still performs better than LASSO and FS, showing the robustness of cLASSO.

In Table B1 in the Supplementary materials we present the results from simulations that considered two toxicity outcomes, T1 and T2. The scenarios were constructed using a subset of the previous efficacy and toxicity models as in Table 1 with toxicity outcome T2 added and constrained at 0.23. The situations considered included a variety of biomarker main effects, dose-biomarker interactions, correlations between the biomarkers and additional noise biomarkers. The results in Table A1 provide similar conclusions regarding the relative merit of cLASSO compared to the other methods as in the single toxicity outcome case.

Table 3.1: Simulation results. Summary of average Efficacy improvement compared with fixed dose with P(Toxicity) constrained to be ≤ 0.2. Results from 1000 simulated trials. Each scenario true logistic models for E and T include main effect for the biomarkers, dose and biomarker-dose interactions, with coefficients as shown below.

Scenarios	Efficacy and Toxicity model coefficients												FS	LASSO	cLASSO	Possible Improvement	cLASSO > LASSO
		Biomarker					Dose	Interactions									
0	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.513	0.518	0.589	0.147	69.0%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
1	E	1	0	0	0	0	1	0	0	0	0	0	0.393	0.659	0.694	0.052	53.3%
	T	-1	0	0	0	0	1	0	0	0	0	0					
2	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.344	0.340	0.375	0.070	56.7%
	T	-1	0	0	0	0	1	.4	.4	.4	-.8	0					
3	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.216	0.332	0.354	0.146	60.4%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
4	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	NA	0.246	0.272	0.146	60.0%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
5	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.692	0.691	0.762	0.148	71.7%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
6	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.433	0.431	0.461	0.148	66.3%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
7	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.561	0.561	0.628	0.147	66.2%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
8	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.378	0.399	0.450	0.200	58.2%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					

Table 3.1 Continued:

Scenarios	Efficacy and Toxicity model coefficients												FS	LASSO	cLASSO	Possible Improvement	cLASSO > LASSO
		Biomarker					Dose	Interactions									
9	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.360	0.402	0.481	0.206	64.6%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
10	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.455	0.466	0.551	0.146	69.3%
	T	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
0⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.618	0.648	0.694	0.132	62.9%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
1⋆	E	1	.2	.3	.1	0	1	0	0	0	0	0	0.410	0.590	0.641	0.055	56.4%
	T	-1	-.2	-.3	-.1	0	1	0	0	0	0	0					
2⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.373	0.497	0.647	0.053	77.5%
	T	-1	-.2	-.3	-.1	0	1	0	.2	-.1	.6	0					
3⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.280	0.429	0.460	0.131	62.1%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
4⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	NA	0.307	0.342	0.133	61.7%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
5⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.788	0.802	0.833	0.134	64.0%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
6⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.512	0.613	0.640	0.133	62.0%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
7⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.671	0.716	0.756	0.134	65.0%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
8⋆	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.391	0.435	0.457	0.153	54.8%

Table 3.1 Continued:

Scenarios	Efficacy and Toxicity model coefficients											FS	LASSO	cLASSO	Possible Improvement	cLASSO > LASSO
		Biomarker					Dose	Interactions								
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
$9^\star$	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.317	0.372	0.435	0.081	65.6%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
$10^\star$	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.519	0.559	0.624	0.132	68.4%
	T	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					

a. The intercept for Efficacy models is 0, for Toxicity models is -1.386.

b. Possible Improvement=$\mathbb{E}^{\mathcal{F}_{Theory}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)$, the percentage of improvement=$\{\mathbb{E}^{\mathcal{F}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}/\{\mathbb{E}^{\mathcal{F}_{Theory}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}$.

c. Scenarios 3, 6, $3^\star$, $6^\star$ have 15 noise covariates with coefficients 0 added; scenarios 4, $4^\star$ have 45 noise covariates with coefficients 0 added.

d. Scenarios 5, 6, $5^\star$, $6^\star$ have doubled sample size of 400.

e. Scenario 7, $7^\star$ have $cor(x_1, x_2, x_3) = 0.6$.

f. Scenarios 8, 9, 10, $8^\star$, $9^\star$, $10^\star$ have mis-specified models.

In scenario 8, $8^\star$, the true effect of covariate x_4 is stepwise at 0, e.g.,logit(P$(E = 1)) = x_1 + d + (0.4x_1 + 0.4x_2 + 0.4x_3 - 0.8I(x_4 > 0))d$.

In scenario 9, $9^\star$, the true models have $exp(x_4)$ as the covariate,e.g.,logit(P$(E = 1)) = x_1 + d + (0.4x_1 + 0.4x_2 + 0.4x_3 - 0.8exp(x_4))d$.

In scenario 10, $10^\star$, an interaction of x_2 and x_3 is included as main effect in both true models for efficacy and toxicity,

e.g.,logit(P$(E = 1)) = x_1 + x_2x_3 + d + (0.4x_1 + 0.4x_2 + 0.4x_3 - 0.8x_4)d$.

3.4 Application

In this section, we applied the proposed method to real data collected from patients with non-small cell lung cancer who received radiation treatment. Patients treated with stereotactic body radiation therapy or with follow-up less than one year were excluded from the analysis, leaving 105 patients in the dataset to be analyzed. Of the 105 patients, 46 had no local, regional or distant progression in two years. Two toxicity outcomes were considered: grade 3+ heart toxicity and grade 3+ lung toxicity. In total, 8 patients had grade 3+ heart toxicity, and 11 patients had grade 3+ lung toxicity that required hospitalization. The clinical features we consider for possible inclusion in models include sex, age, current smoker, Karnofsky Performance Status (KPS), concurrent chemotherapy, simple stage, T-stage, N-stage of the cancer, as shown in Table.3.2. We also include pre-treatment cytokines level such as interferon γ (IFN-γ), interleukin-1 β (IL-1β), interleukin-2 (IL-2), interleukin-6 (IL-6), tumor necrosis factor α (TNF-α) as prognostic factors. Patients in this study received different doses ranging from 45 Gy to 96 Gy, partially due to the preference of different clinicians as well as the stage of the disease, location of the tumor and the patients performance status. The dose to the tumor site (efficacy dose) is different from the dose received by the lung and heart, but we assume the ratio of them is fixed for each patient. When the optimal efficacy dose is chosen within the observed dose range (45 - 96 Gy), it is multiplied by this known fixed (for each patient) ratio to obtain the lung and heart dose corresponding to the selected tumor dose. There are 14 patients with no cytokine data collected, and multiple imputation with all the covariates and outcomes included is applied to fill in the missing values.

For the given set of doses in the study, the average probability of no progression in two years (efficacy) is 0.438, the average probability of heart toxicity is 0.076, the average probability of lung toxicity is 0.105, and average tumor dose across patients is 71.20 Gy. The goal of this analysis is to estimate an optimal dosing rule that

Table 3.2: Descriptive statistics of patients (n=105)

Variable	Mean	Range
Age (Years)	65.43	39.60 - 85.20
KPS	85.52	60 -100
IFN-γ	113.31	0.52 - 6547.50
IL-1β	10.26	0.04 - 92.61
IL-2	23.50	0.04 - 312.22
IL-6	41.93	0.07 - 730.84
TNF-α	18.48	0.54 - 149.37
Tumor dose (Gy)	71.20	45.66 - 96.08
Lung dose (Gy)	14.47	3.17 - 26.11
Heart dose (Gy)	12.22	0.02 - 46.13
Variable	Category	Percentage
Gender	Female	24
	Male	76
Smoking	Current	42
	Never or former	58
chemotherapy	Yes	85
	No	15
Simple stage	1	10
	2	10
	3	79
	4	1
T stage	1	18
	2	23
	3	27
	4	32
N stage	0	23
	1	12
	2	45
	3	20

Table 3.3: Variable selections in each model

Method		cLASSO		
Model		Efficacy	Heart toxicity	Lung toxicity
Main effect	Dose	$\star$	$\star$	$\star$
	Age			$\star$
	Sex			
	Smoking			
	KPS	$\star$	$\star$	
	Chemotherapy			$\star$
	S stage			
	T stage		$\star$	
	N stage	$\star$		
	IFN-γ	—	—	
	IL-1β	—	—	
	IL-2	—	—	
	IL-6	—	—	
	TNF-α	—	—	$\star$
Interactions with dose	Age	$\star$		$\star$
	Sex			$\star$
	Smoking			
	KPS			
	Chemotherapy			
	S stage			
	T stage			
	N stage		$\star$	
	IFN-γ	—	—	
	IL-1β	—	—	$\star$
	IL-2	—	—	
	IL-6	—	—	
	TNF-α	—	—	
Estimated outcomes if these patients were treated at optimal doses		0.485	0.077	0.108

$-$ represents a covariate which is not considered for inclusion in the model.
$\star$ represents a covariate selected by cLASSO for the corresponding model.
Empty cell represent covariates considered for inclusion but not selected by cLASSO.

maximize the probability of no progression in 2 years, with heart and lung toxicity level no greater than observed overall toxicity for this population of patients. The efficacy model for the probability of no progression in two years including as covariates the 8 clinical features with their interactions with tumor dose as well as tumor dose has in total 17 possible covariates. The heart toxicity model is built similarly, and the lung toxicity model also includes the 5 most important cytokines and their interaction with lung dose. Table 3.3 showed the covariates selection by cLASSO in each model. The random walk method of selecting θ ran for 5000 iterations to ensure convergence. With the models built by cLASSO, using the selected optimal dose for each patient gave an expected efficacy of 0.485, an expected heart toxicity at 0.077, an expected lung toxicity at 0.108, and the average tumor dose across patients was 80.21 Gy. With similar expected lung toxicity and heart toxicity rates, the average efficacy increased by 0.047 from 0.438 to 0.485, which is a clinically significant improvement. Fig 3.5 showed the selected optimal dose for each patient. Some patients were recommended lower doses comparing to their actual doses received, while more patients were recommended higher doses. Mean optimal tumor dose is 80.2 Gy, which is higher than mean actual tumor dose 71.2 Gy.

3.5 Discussion

In this paper, we propose an optimal individualized dose finding rule by maximizing utility functions for individual patients. This approach maximizes overall efficacy at a prespecified constraint on overall toxicity. We model the binary efficacy and toxicity outcomes using logistic regression with dose, biomarkers and dose-biomarker interactions. To incorporate the larger number of biomarkers and their interaction with doses, we employed the LASSO with linear constraints on the dose related coefficients to constrain the dose effect to be non-negative. Simulation studies show that this approach can improve efficacy without increasing toxicity relative to fixed

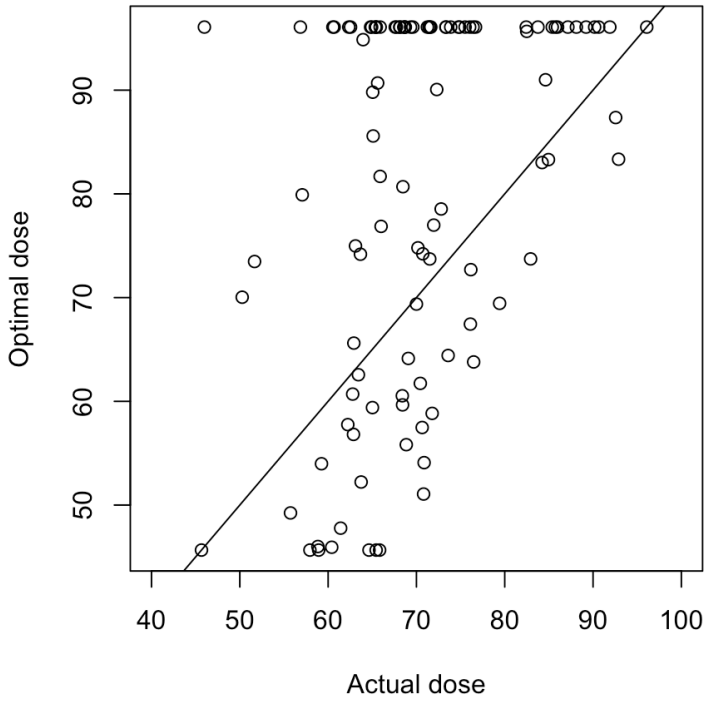

Figure 3.5: Comparison of selected optimal dose and actual dose received for all the patients.

dosing. Constraining each patient's estimated dose-efficacy and dose-toxicity curves to be non-decreasing improved performance relative to standard LASSO. This utility method was extended to multiple toxicities.

To force the dose-toxicity or dose-efficacy curve to be non-decreasing with dose, the constraints for linear combination of dose related coefficients only ensure that the patients in the current data satisfy this monotonicity criteria, but monotonicity is not guaranteed for all future patients whose x is not in the observed data. An alternative approach that would ensure monotonicity with respect to dose for all patients would be to force all relevant dose and dose-covariate coefficients to be non-negative. But with the dose-biomarker interactions, it is unnecessary to force all the dose related coefficients to be non-negative, because some of them could be negative but the linear combination of them is non-negative for a selected range of dose and covariate values. Thus, simply constraining all the dose related coefficients to be non-negative is too

conservative. It is noted that our method is not appropriate in cases where the toxicity or efficacy endpoint may first increase and then decrease with increasing dose, but is still applicable when there is an increasing effect followed by a plateau.

While we implemented a constrained version of LASSO, other penalized regression approaches such as Elastic Net could also be considered [56], or Bayesian methods using Bayesian LASSO [36], or other Bayesian variable selection methods such as "spike-and-slab" [22].

Our method constrains the population averaged toxicity level to be below a given tolerance level. This does not explicitly put any upper bound on the expected toxicity probability for an individual patient. Our method could be modified by including an upper bound on the probability of toxicity for each patient. Thus, in addition to constraints on average toxicity, we also consider adding constraints to individual toxicity, i.e., adding large penalty for extremely high toxicity, which will make the utility function more complex. An indirect way of achieving this would be to use a non-linear function of the probability of toxicity, rather than just the toxicity rate in equation (3.1).

In this paper we have considered binary outcomes and logistic models that included main effect and dose-biomarker interactions. The method could be generalized to other type of outcomes, such as censored survival times for the efficacy outcome. An ordinal outcome for toxicity could also be accommodated by requiring a different tolerance threshold for each level of toxicity. More flexible forms for the effect of dose and biomarkers could also be considered (e.g., regression splines), and provided the dose monotonicity constraint can be algebraically formulated, the cLASSO would still be applicable.

CHAPTER IV

Utility Based Approach in Individualized Optimal Dose Selection Using Machine Learning Methods

4.1 Introduction

Precision medicine is an approach for the treatment of a disease that takes into account individual variability. In many situations, the clinical decision is whether to give a particular treatment or a standard treatment to the patient, and the statistical goal is to identify the subgroup of patients likely to derive benefit from the treatment compared to standard treatment. In other settings, the treatment choice is not binary and in particular consists of choosing the value of a continuous variable, which we call dose. The goal in this setting is to find the optimal dose for each patient to maximize the benefit of treatment. With the development of biomarkers, the goal can be achieved by evaluating how the biomarkers moderate the treatment effect on the outcome or outcomes. In statistical terms moderation of the treatment effect can be formulated as the interaction between the biomarkers and dose in a model.

In this paper we consider the setting where there is an existing dataset from previously treated patients that contains treatment dose, observed efficacy and toxicity outcomes and covariates possibly including biomarkers. We further assume there is heterogeneity in the treatment dose given in this dataset. The statistical goal is to

analyze the data to learn an individualized dosing rule giving an optimal dose as a function of patient level covariates.

Supervised learning in the form of regression (for continuous outputs) and classification (for discrete outputs) is an important constituent of statistics and machine learning. Widely used parametric models such as linear regression and logistic regression are simple, but can suffer from model mis-specification. Concerns about mis-specification can be reduced by including extra features, such as interactions and splines, or there are many nonparametric methods that can be used instead. Because of their greater flexibility, there is less concern about model mis-specification, but also increased potential for overfitting. Decision Trees such as Classification And Regression Trees (CART) and random forest, built upon CART, are easy to understand and interpret and have been used in determining the optimal treatment [12]. Their tree structure handles the interaction well but also increases the potential of overfitting. Methods based on kernel machines are also flexible and have good performance with high-dimensional data. Examples of these are Support Vector Machines as used in the outcome weighted learning(OWL) [55] method, and Gaussian process [51]. Modern Bayesian semiparametric and nonparametric models also provide posterior uncertainty quantification which can be used to provide uncertainty estimates of patient specific treatment decisions. An example of this is using Bayesian Additive Regression Trees (BART) to provide individual treatment rules (ITR). This approach provides the uncertainty of the outcome associated with the optimal ITR [28].

In early phase studies in oncology, it is typical to describe the patient outcome in terms of both toxicity (T) and efficacy (E). The benefit for each patient can be defined based on combining these two outcomes. By maximizing the potential benefit for each patient, the individualized optimal treatment or dose will be selected. Several strategies have been proposed for identifying an optimal treatment or dose based on the trade-off between efficacy and toxicity. To achieve maximum efficacy with tolera-

ble toxicity, a utility function can be used as a weighted difference of probabilities of efficacy and toxicity [23]. More generally, a utility matrix is elicited from clinicians by assigning numerical utilities to each possible efficacy and toxicity outcome pair, which allows different preferences for different outcomes and then the utility function, defined as the expected value of the utility, at different dose levels are compared [19]. Contours characterizing the trade-off between E and T is an alternative and flexible approach. For this, the set of pairs for the probability of E and the probability of T on the same contour are equally desirable and the dose that maximizes the desirability is selected [46].

In oncology and other disease settings, it is frequently reasonable to assume that increasing dose leads to increased toxicity and efficacy. However, the increase may not be strictly monotone over the whole range of possible doses. For example, FDA guidance on cellular and gene therapy products mentions that indicators of potential benefit appear to plateau above a certain dose [11]. Thus, we will consider constraining the estimated dose-efficacy and dose-toxicity relationship to be non-decreasing for all patients. The effect of this should be to reduce the estimation noise, improve efficiency and hence yield more reliable results.

In previous work we built separate models for E and T and did not explicitly consider correlation between E and T [23]. In this paper we relax the assumption of independence and use flexible machine learning methods to build the models for E and T using dose and biomarkers. In section 3, we propose the use of random forest and Gaussian process models to build marginal models and link them with a Gaussian copula. We also build models on joint probabilities and use PAVA isotonic transformation to give a non-decreasing dose-efficacy and dose-toxicity relationship. In section 4, we consider alternative uses of the utility function to select an optimal dose, including the addition of constraints with clinical motivation as well as the uncertainty of the utility function. A simulation study to compare different model

building and optimal dose finding methods under several scenarios is summarized in Section 5. In section 6, we illustrate the proposed methods in a dataset of patients with non-small lung cancer treated with radiation therapy. We close with discussion in section 7.

4.2 Utility function and matrix

We assume the available data, $\mathcal{D}_i = (x_i, d_i, E_i, T_i), i = 1, ...N$, comprises N independent and identically distributed copies of (x, d, E, T), where x is a Q-dimensional vector of subject-specific features, $d \in [-1, 1]$ denotes continuous dose of treatment, E is the binary efficacy outcome, and T is the binary toxicity outcome. Denote $p_E(d, x) = \Pr(E = 1|d, x)$ and $p_T(d, x) = \Pr(T = 1|d, x)$ as the marginal probabilities for the efficacy and toxicity outcomes. One way to combine the efficacy and toxicity outcomes is through a utility function such as $U(\boldsymbol{p}(d, x), \theta) = p_E(d, x) - \theta p_T(d, x)$ with $\theta > 0$. θ can be pre-specified by physicians or calculated as a tuning parameter to meet a pre-specified level of toxicity in the population [23]. Alternatively, a unique utility value can also be specified for each possible bi-variate patient outcome (E, T) in a utility matrix, as shown in Table 4.1. We assign U_{10} the highest value to the best possible outcome (efficacy and no toxicity), and U_{01} the lowest value to the worse possible outcome (toxicity and no efficacy). U_{00} and U_{11} have values in between and can be larger or smaller or equal to each other. Denote the joint probability of E and T given d and x as in Table 4.2. We define a utility function as the expectation of this utility matrix with respect to the joint probability of E and T at dose d and covariates x. The goal is then to maximize the utility function, or the expectation of the utility matrix across dose for a fixed x.

The utility function above, $U(\boldsymbol{p}(d, x), \theta) = p_E(d, x) - \theta p_T(d, x)$, corresponds to one with $(U_{00}, U_{10}, U_{01}, U_{11}) = (0, 1, -\theta, 1 - \theta)$. More generally, without loss of generality, we can assign the highest utility as 1 and lowest utility as 0. Then we consider the

Table 4.1: Utility matrix

	E=0	E=1
T=0	U_{00}	U_{10}
T=1	U_{01}	U_{11}

Table 4.2: Table of probabilities

	E=0	E=1	Row Sum
T=0	p_{00}	p_{10}	$1 - p_T$
T=1	p_{01}	p_{11}	p_T
Column Sum	$1 - p_E$	p_E	1

utility matrix with $0 < \omega_1 < 1$, $0 < \omega_2 < 1$ that are pre-specified, as shown in Table 4.3.

Table 4.3: utility matrix with two parameters

	E=0	E=1
T=0	ω_1	1
T=1	0	ω_2

We define the Utility function $\bar{U}(\boldsymbol{p}(d,x), \boldsymbol{\omega})$ as the expectation of the utility matrix in Table 4.3,

$$\begin{aligned} \bar{U}(\boldsymbol{p}(d,x), \boldsymbol{\omega}) &= \omega_1 p_{00}(d,x) + p_{10}(d,x) + \omega_2 p_{11}(d,x) \\ &= \omega_1 + (1-\omega_1)p_E(d,x) - \omega_1 p_T(d,x) + (\omega_1 + \omega_2 - 1)p_{11}(d,x) \end{aligned} \tag{4.1}$$

When $\omega_1 + \omega_2 = 1$, $\bar{U}(\boldsymbol{p}(d,x), \boldsymbol{\omega}) = \omega_1 + (1-\omega_1)p_E(d,x) - \omega_1 p_T(d,x)$, which yields equivalent solution to $U(\boldsymbol{p}(d,x), \theta) = p_E(d,x) - \theta p_T(d,x)$. The gain in utility when moving from $E = 0$ to $E = 1$ is the same for both levels of T ($1 - \omega_1 = \omega_2$). For this special case the models for E & T can be built separately [23], and no assumption concerning independence or dependence of E and T is needed. However, when $\omega_1 + \omega_2 \neq 1$, this is not the case and the utility function does depend on the correlation between efficacy and toxicity.

There are various approaches to modeling the joint distribution of E and T of

which we consider three: model the marginal distributions for E and T and then assume independence, or use a copula to link them as the joint distribution, or directly model the 4 level categorical outcome (E, T) on d and x.

4.3 Model building

Instead of using parametric models, we will consider using non-parametric approaches (e.g. random forest) and kernel machines (e.g. Gaussian process) to model the joint distribution of E and T.

4.3.1 Random forest

First, we use random forest (RF) to build models for E and T separately to obtain estimates of $\hat{p}_E(d, x)$ and $\hat{p}_T(d, x)$ for each subject. Here we use Out Of Bag (OOB) predictions to avoid the overfitting. After obtaining estimated marginal probabilities, we can assume independence of efficacy and toxicity so that the joint distribution of (E, T) is given by the product of the marginals. Alternatively, we can use a copula to link the marginals as described in the appendix C.1. We can plug the $\hat{p}_E(d, x)$ and $\hat{p}_T(d, x)$ for each subject into the copula function and maximize the loglikelihood across the dataset $\mathcal{D}_N$ to estimate $\hat{\alpha}$, the correlation parameter in the copula. *Rborist* by Seligman was used to implement the random forest due to its computational efficiency [42].

In random forest, the predicted probabilities of E or T may decrease over some range of dose for some patients. To prevent this, we can apply constraints during the estimation of the random forest. Specifically, whenever a split on dose is selected, we require the left node (lower dose) to be associated with lower probabilities of E or T than the right node (larger dose) [42].

Instead of modeling marginals and linking them to obtain the joint distribution, we can model the joint probabilities directly. When doing this using methods such

as random forests, we can't add monotonicity directly in the model since the desired monotonicity is in terms of each marginal and not in the four joint probabilities $p_{00}, p_{10}, p_{01}, p_{11}$. Thus in this setting we use a post estimation procedure in which we adjust $\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11}$ to obtain $\hat{p}^*_{00}, \hat{p}^*_{10}, \hat{p}^*_{01}, \hat{p}^*_{11}$ so that the corresponding p_E and p_T are monotone in d for all x. Specifically, we use the Pool-Adjacent-Violators Algorithm (PAVA) isotonic transformation for the marginal distribution, then we will find a set $(\hat{p}^*_{00}, \hat{p}^*_{10}, \hat{p}^*_{01}, \hat{p}^*_{11}\ )$ which is close to $(\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11})$ and also satisfies the monotonicity constraint.

For every patient, using the model of categorical outcomes built on the dataset, we have predictions of $(\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11})$ for all the dose levels and thus $\hat{p}_E$ and $\hat{p}_T$. Then we apply PAVA on $\hat{p}_E$ and $\hat{p}_T$ to get $\hat{p}^*_E$ and $\hat{p}^*_T$ which is non-decreasing in dose across the feasible dose range. At each dose, the joint probability $\hat{p}^*_{11}$ has a range of $[max(0, \hat{p}^*_E + \hat{p}^*_T - 1), min(\hat{p}^*_E, \hat{p}^*_T)]$. One way to perform the adjustment is to minimize $|\hat{p}^*_{11} - \hat{p}_{11}|$ within the allowed range to get $\hat{p}^*_{11}$, then $\hat{p}^*_{10}, \hat{p}^*_{01}, \hat{p}^*_{00}$ can be calculated sequentially from the known $\hat{p}^*_{11}, \hat{p}^*_E, \hat{p}^*_T$. Another way is to minimize $|\hat{p}^*_{11} - \hat{p}_{11}| + |\hat{p}^*_{10} - \hat{p}_{10}| + |\hat{p}^*_{01} - \hat{p}_{01}| + |\hat{p}^*_{00} - \hat{p}_{00}|$, which involves all the joint probabilities.

4.3.2 Gaussian process

As a second approach to estimating the marginal distributions, we use a Gaussian process (GP) approach. Specifically, we model the outcome $\boldsymbol{y}$ as a distorted version of the process $\boldsymbol{f}$, where $\boldsymbol{f}$ is a non-linear latent function and is assigned a Gaussian process prior. For example, a continuous outcome might use a normal distribution around $\boldsymbol{f}$, while a binary outcome would use a logistic link or probit link with $\boldsymbol{f}$. With the Gaussian prior

$$\boldsymbol{f} \sim GP(\boldsymbol{0}, \boldsymbol{K}) = p(\boldsymbol{f}),$$

and likelihood

$$\boldsymbol{y} \sim \prod_{i=1}^{n} p(y_i|f_i),$$

the Gaussian posterior of $p(\boldsymbol{f}|\boldsymbol{y})$ is given by

$$p(\boldsymbol{f}|\boldsymbol{y}) \sim \prod_{i=1}^{n} p(y_i|f_i) N(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}).$$

where $\boldsymbol{K}$ is a kernel function that represents the distance between subjects in the covariate space. Multiple kernel functions could be used and we use a Gaussian kernel with each element in $\boldsymbol{K}$ defined by

$$k(x_i, x_j) = \eta^2 exp(-\frac{1}{2}\sum_{q=1}^{Q} \rho_q^2 (x_i^{(q)} - x_j^{(q)})^2)$$

for subjects i and $j = 1, ..., N$ and dimension $q = 1, ..., Q$. η^2 is called the signal variance or the magnitude, and ρ_q's are the characteristic length-scales of the input-space for each dimension of x. This covariance function implements automatic relevance determination (ARD) [34], since the length-scale ρ_q determines how relevant an input is: if the length-scale is small, the covariance will become almost independent of that input, effectively removing it from the inference. The ARD can be used in selection of the dimension x related to the outcome, especially with high dimension data or sparse data. We assign priors to η, ρ_q and estimate the posterior distribution of them from the data.

In our setting, we assume $\boldsymbol{f}_E \sim N(0, \boldsymbol{K}_E), \boldsymbol{f}_T \sim N(0, \boldsymbol{K}_T)$ where $\boldsymbol{K}_E, \boldsymbol{K}_T$ are different kernel functions for E, T. For each subject, the marginal probabilities are,

$$\Pr(E = 0) = 1 - p_E = \Phi(f_E + a), \Pr(T = 0) = 1 - p_T = \Phi(f_T + b),$$

here we use a probit link to connect f to the outcome. Because f are centered around

0, intercepts a, b are used for the unbalanced outcomes. Then the joint probability can be obtained by the Gaussian copula

$$p_{00}(d,x) = \Phi_2[\Phi^{-1}(1-p_E), \Phi^{-1}(1-p_T)|\alpha] = \Phi_2[f_E + a, f_T + b|\alpha]$$

Alternatively, when assume independence of efficacy and toxicity,

$$p_{00}(d,x) = \Phi(f_E + a)\Phi(f_T + b)$$

The outcome (E, T) for each subject follows $Multinomial(p_{00}, p_{10}, p_{01}, p_{11})$ and we can use a Bayesian approach to estimate $(\eta_E, \rho_{qE}, \eta_T, \rho_{qT}, a, b, \alpha)$ jointly with some prior distribution. *rstan* was used to implement the Bayesian analysis of Gaussian process since it is more efficient and explores complicated posterior distributions better compared to *rjags* [45].

For Gaussian process, there are no methods to ensure monotonicity of predictions with respect to dose across the entire covariate space. Rather, virtual data points can be included and the derivative of the Gaussian process (with respect to dose) at those points can be forced to be non-negative during estimation [39].

4.4 Optimal dose finding

For a subject with covariates x, we can directly calculate the optimal dose when the true probabilities $\boldsymbol{p}(d,x)$ are known, by maximizing the utility function across possible dose values. That is

$$d_{opt}(x) = argmax \quad \bar{U}(\boldsymbol{p}(d,x), \boldsymbol{\omega}) \tag{4.2}$$

4.4.1 Modified utility functions

In oncology clinical studies, efficacy and toxicity usually increase with dose simultaneously, thus high toxicity could be observed with high efficacy. So when maximizing efficacy, it is usual to have a pre-specified toxicity limit that needs to be satisfied. We can achieve this by modifying the utility function by increasing the penalty on higher values of $p_T(d, x)$. As an illustration we will modify the utility function such that if $p_T(d, x) \geq 0.3$, the weight for toxicity will be tripled, giving a utility function of

$$\bar{U}_1(\boldsymbol{p}(d, x), \boldsymbol{\omega}) = \bar{U}(\boldsymbol{p}(d, x), \boldsymbol{\omega}) - 2\omega_1 p_T(d, x)\mathbb{I}\big[p_T(d, x) \geq 0.3\big]. \tag{4.3}$$

The individual optimal dose is obtained by maximizing the utility function, and this optimal dose differs across patients. It is possible that the optimal dose is at the boundary of the feasible dose region and far from a population fixed dose d_{fix}, that is the same for everyone. The d_{fix} can be pre-specified as a standard value, or estimated by fitting a dose-only model without covariates x in it. It might be clinically desirable to avoid extreme doses. This can be achieved by shrinking the individual optimal dose towards the population fixed dose, by including a penalty in the utility function.

$$\bar{U}_2(\boldsymbol{p}(d, x), \boldsymbol{\omega}) = \bar{U}(\boldsymbol{p}(d, x), \boldsymbol{\omega}) - \delta|d - d_{fix}| \tag{4.4}$$

$$\bar{U}_3(\boldsymbol{p}(d, x), \boldsymbol{\omega}) = \bar{U}(\boldsymbol{p}(d, x), \boldsymbol{\omega}) - \delta(d - d_{fix})^2 \tag{4.5}$$

where δ is the penalty parameter and d_{fix} is the population fixed dose.

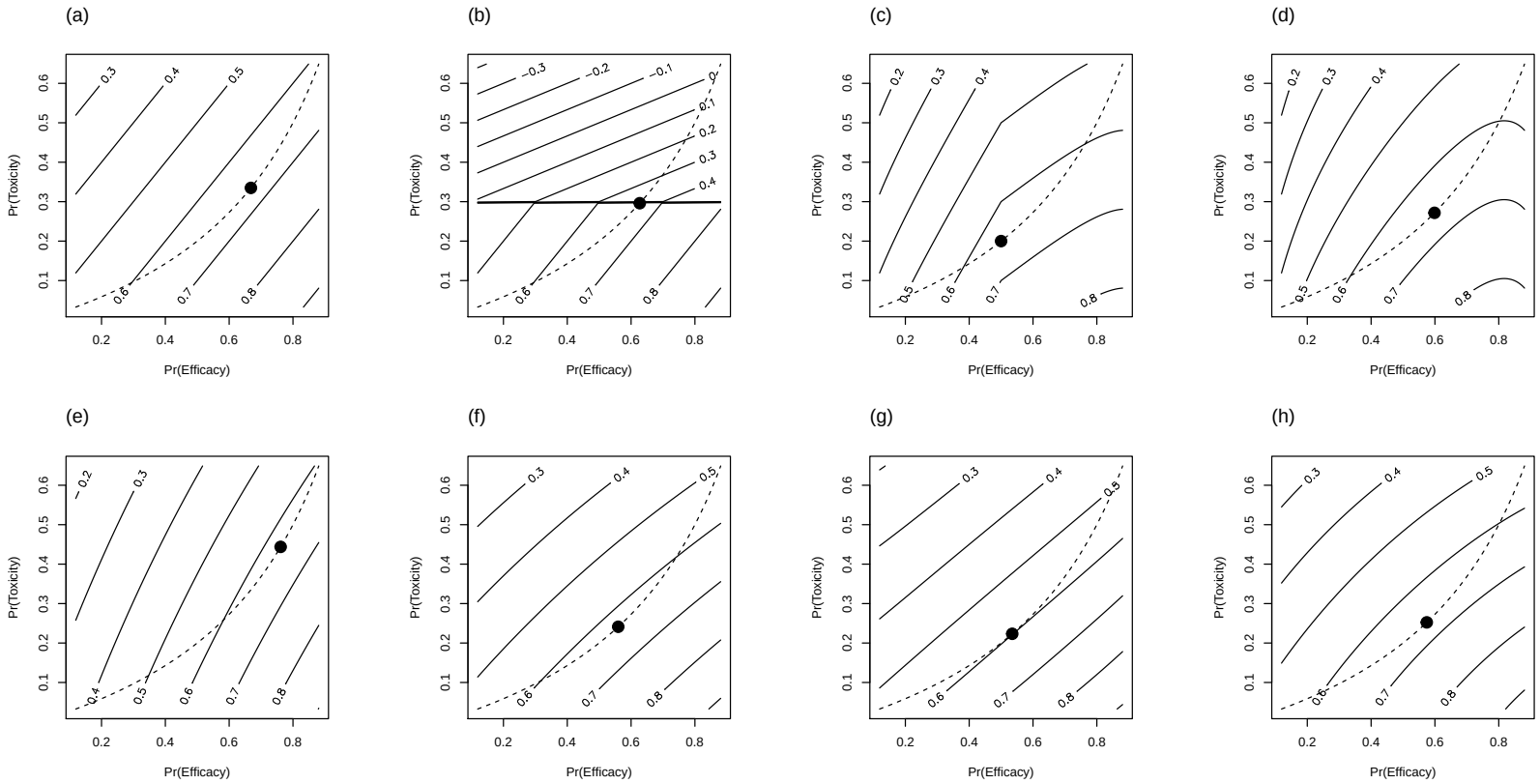

Figure 4.1: Comparison of utility function contours. (a) $\bar{U}$ with $\omega_1 = \omega_2 = 0.5$, (b) $\bar{U}_1$ with $\omega_1 = \omega_2 = 0.5$, (c) $\bar{U}_2$ with $\omega_1 = \omega_2 = 0.5$, $\delta = 0.1$, (d) $\bar{U}_3$ with $\omega_1 = \omega_2 = 0.5, \delta = 0.1$, (e) $\bar{U}$ with $\omega_1 = 0.3, \omega_2 = 0.5$ and independent E and T, (f) $\bar{U}$ with $\omega_1 = 0.5, \omega_2 = 0.3$ and independent E and T, (g) $\bar{U}$ with $\omega_1 = 0.5, \omega_2 = 0.3$ and cor(E,T)=0.8, (h) $\bar{U}$ with $\omega_1 = 0.5, \omega_2 = 0.3$ and cor(E,T)=-0.8.

Figure 4.1 shows the trade-off contours over the two-dimensional space with different utility functions. The pairs of (p_E, p_T) on each contour are considered to have the same utility function, as denoted by the solid lines. For a given patient with covariates x, the $p_E(d, x)$ and $p_E(d, x)$ among the possible dose range is shown as a dashed line, and the dose that maximize the utility function is selected, as shown by the points. When $\omega_1 + \omega_2 = 1$, the utility function is a weighted difference of p_E and p_T and results in straight lines. All the other utility functions have curved or broken lines. For $\bar{U}_1$, the utility function decreases rapidly when $p_T(d, x)$ is above the pre-specified toxicity limit of 0.3. $\bar{U}_2$ and $\bar{U}_3$ have different penalties for dose with the l_2-penalty giving smoother contours. When $\omega_1 + \omega_2 \neq 1$, the four plots (e) to (h) indicate the influence of different ω_1, ω_2 values and the correlations of E and T. The lines are curved with $p_{11}(d, x)$, and the joint distribution of E and T plays a role. These plots shows different choices of optimal dose as the utility function changes. For (f) to (h), the correlations are 0, 0.8 and -0.8 with the same set of ω_1, ω_2 and marginal probabilities for a given patient, the utilities differs a lot, but the optimal doses are similar.

4.4.2 Use uncertainty of the estimation

The random forest will give the point estimate of the probabilities for the possible doses, which is $\hat{\boldsymbol{p}}(d, x)$ and can be plugged into equation (4.2). Then the individual optimal dose for subject with covariates x is the dose that maximizes the weighted sum of the utility values using the joint probability estimates. When using Bayesian estimation, we will get the posterior distribution of the joint probabilities $\boldsymbol{p}(d, x)$ for the possible doses, and thus the posterior distribution of $\bar{U}(\boldsymbol{p}(d, x), \boldsymbol{\omega})$. The optimal dose for subject with covariates x is selected as the dose that maximize the posterior mean of $\bar{U}(\boldsymbol{p}(d, x), \boldsymbol{\omega})$ given the data $\mathcal{D}_N$.

$$d_{opt}(x) = argmax \quad E(\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega})|\mathcal{D}_N) \tag{4.6}$$

To assess the uncertainty of the estimation, we can consider the posterior distribution of the utility function. Then an alternative way to define the optimal dose for subject with covariates x is as the dose that maximizes the posterior probability that the utility function is higher than the utility function at some fixed dose which is the same for everyone.

$$d_{opt} = argmax \quad P(\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega}) > \bar{U}(\boldsymbol{p}(d_{fix},x),\boldsymbol{\omega})|\mathcal{D}_N) \tag{4.7}$$

4.5 Simulation studies

4.5.1 Settings and scenarios

To evaluate the performance of the different methods, we conducted a simulation study. We consider 11 scenarios by varying the true parameter coefficients, the functional form (e.g. exponential or binary or linear), the number of biomarkers, and the sample size. For each scenario and each method of analysis and definition of optimal dose, we fit the models using the training dataset, and evaluated their performance under the true models which are known only in this simulation setting.

In scenario 1, we generated N=200 observations in a training dataset by simulating 5 i.i.d covariates, $x_1, ...x_5$ from a standard normal distribution, d from Uniform(-1,1), and binary outcomes E and T from the regression model described below with a copula to link them. The marginal probabilities are defined by models

$$p_E(d,x) = \Phi(\beta_{0,E} + (x,d,dx)\beta_E), \qquad p_T(d,x) = \Phi(\beta_{0,T} + (x,d,dx)\beta_T)$$

and $\alpha = 0.8$ was used in the Gaussian copula to link the two marginal distribu-

tions. In scenario 1, $(\beta_{0,E}, \beta_E) = (0, 0.49, -1.11, 0.77, 1.51, 0, 1, 0.23, 0.61, 0, 1.69, 0.5)$, $(\beta_{0,T}, \beta_T) = (-1.386, 0, 1.14, -0.33, 0, 0, 1, 0.03, 0.6, 0, -0.42, 1.04)$. In generating x's, we also applied the constraints that x must satisfy $1+0.23x_1+0.61x_2+1.69x_4+0.5x_5 > 0$ and $1+0.03x_1+0.6x_2-0.42x_4+1.04x_5 > 0$ to reflect the non-decreasing dose-efficacy and dose-toxicity curves. This excludes up to 40% of simulated observations.

The other scenarios considered are shown below in Table 4.4 . In all scenarios the true model coefficients were chosen so that the proportion of observations with $E = 1$ was about $50 - 70\%$ and with $T = 1$ was about $10 - 30\%$.

Table 4.4: List of scenarios

S0	The true E & T model has only d without covariates x
S1	The true E & T models have x, d, dx
S2	The true E & T models have x, d, no dose-covariate interactions
S3	S1 with x_1, x_2, x_3 correlated with coefficient 0.6
S4	S1 with 15 noise covariates $x_6, ..., x_{20}$ added to the data
S5	S1 with 195 noise covariates $x_6, ..., x_{200}$ added to the data
S6	S1 with sample size = 400
S7	S1 with binary covariates x
S8	The true E & T model has x, d, dx, as well as covariate interactions xx
S9	The true models have $exp(x_4)$ as the covariate
S10	The true models have $I(x_4 > 0)$ as the covariate

The simulation consist of the following steps iterated 1000 times. 1) Generate a training dataset of the specified size (N=200 or 400). 2) Fit models for efficacy and toxicity on the training data. 3) Generate a validation dataset following the same approach as for the training data. 4) Calculate the estimated optimal dose for each patient in the validation data using the models estimated from the training data and a particular method for selecting optimal dose based on the models. Also calculate the true efficacy and toxicity probabilities for each patient, conditional on the estimated optimal dose, using the true models. 5) Calculate the average and standard deviation of the estimated optimal dose values, average probability of efficacy, average probability of toxicity, average value of the true utility function, and average improve-

ment in utility defined as the fraction of the difference between the expectation of the utility from the true model and the expectation of the utility from using a fixed dose. The metrics in step (5) are then averaged over the 1000 simulations to provide a comparison of the various methods for model building and optimal dose selection across various scenarios.

We consider several variations of random forest modeling. To allow for correlated efficacy and toxicity outcomes we take three approaches consisting of using a copula method to link the marginal models, assuming conditional independence and also directly modeling the 4 level bi-variate outcome. We also implement a version of random forest in which the dose effects on efficacy and toxicity are constrained to be non-negative for all patients and compare these results to unconstrained estimation. When using random forest to model the 4 level categorical outcome, we also assess the impact of a post-estimation use of a PAVA algorithm to enforce monotonicity.

For the Gaussian process, we compared different priors for ρ_q including inverse gamma prior, Laplace prior and horseshoe prior. They have similar performance with low dimension of covariates, but the horseshoe prior is better with high dimension as it allows variable selection. Specifically, $\rho_q \sim N(0, \lambda_q^2\tau^2)$ for $q = 1, ..., Q$, where $\lambda_q \sim cauchy(0, 1), \tau \sim cauchy(0, 1)$. The global parameter τ pulls all the weights globally towards zero, while the thick half-Cauchy tails for the local scale λ_q allow some of the weights to escape the shrinkage. With large τ, all ρ's have diffuse priors with very little shrinkage toward 0, but small τ will shrink all the ρ's to 0. All the other parameters have a standard normal distribution as priors. For Gaussian process, we tried both using copula and assuming independence to link the two marginals. The results were very similar so we only include those corresponding to the independence assumption. Although we did investigate imposing monotonicity constraints in Gaussian process modeling through the use of virtual data points, we did not include these methods in our simulation because they are time consuming and in limited simulations appeared

to have little effect on the results. All the methods considered are summarized in Table 4.5.

The utility matrix plays an important role in optimal dose selection so for each scenario, we consider three utility matrices. The first places higher utility on the outcome of both efficacy and toxicity compared to neither efficacy nor toxicity ($\omega_1 = 0.3, \omega_2 = 0.5$). The corresponding utility function is $\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega}) = 0.3+0.7p_E(d,x)-0.3p_T(d,x)-0.2p_{11}(d,x)$. The second utility matrix we consider, weights these two possible outcomes equally ($\omega_1 = \omega_2 = 0.5$). In this setting the correlation plays no role as it cancels out of the utility function, as $\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega}) = 0.5 + 0.5p_E(d,x) - 0.5p_T(d,x)$. In the third utility matrix we consider, the outcome of no efficacy and no toxicity is preferred to the outcome of both ($\omega_1 = 0.5, \omega_2 = 0.3$). The utility function is $\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega}) = 0.5 + 0.5p_E(d,x) - 0.5p_T(d,x) - 0.2p_{11}(d,x)$.

Table 4.5: List of methods. RF denotes random forest, GP denotes Gaussian process

	Model building	Optimal dose selection
M1	True model	$\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega})$
M2	RF on marginals with copula	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega})$
M3	RF on marginals with independence	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega})$
M4	RF on marginals monotone on d with copula	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega})$
M5	RF on marginals monotone on d with independence	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega})$
M6	RF on marginals monotone on d with copula	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega}) - 2\omega_1\hat{p}_T(d,x)\mathbb{I}\big[\hat{p}_T(d,x) \geq 0.3\big]$
M7	RF on marginals monotone on d with copula	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega}) - \delta(d-d_{fix})^2$
M8	RF on categorical outcome	$\bar{U}(\hat{\boldsymbol{p}}(d,x),\boldsymbol{\omega})$
M9	RF on categorical outcome and PAVA to minimize $\lvert\hat{p}^*_{11} - \hat{p}_{11}\rvert$	$\bar{U}(\hat{\boldsymbol{p}^*}(d,x),\boldsymbol{\omega})$
M10	GP with independence assumption	$P(\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega}) > \bar{U}(\boldsymbol{p}(d_{fix},x),\boldsymbol{\omega})\lvert\mathcal{D}_N)$
M11	GP with independence assumption	$E(\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega})\lvert\mathcal{D}_N)$
M12	Fixed dosing: multinomial logistic regression on dose	$\bar{U}(\hat{\boldsymbol{p}}(d),\boldsymbol{\omega})$

4.5.2 Results

Figure 4.2 shows the scatter plot of the estimated optimal doses for one representative dataset with $n = 200$ under scenario 1 with $\omega_1 = 0.3, \omega_2 = 0.5$. Under the true model (M1), the optimal dose would be the minimum or maximum possible value for many patients. The optimal doses selected by different methods are compared with the true optimal dose. For the random forest models on the marginals (M2, M3, M4, M5), the optimal doses for many patients are near the optimal fixed dose, which is 0.63 for this dataset. There is little difference between methods using copula and assuming independence (M2 vs M3, M4 vs M5), suggesting that use of the copula does not have a large impact on the distribution of selected doses. As expected, method M6 tends to avoid maximum dose and select lower doses than method M4 due to the individual toxicity limit in the utility function. For method M7, the majority of the patients have optimal doses around the fixed dose because of the dose penalty in the utility function. For the random forest with categorical outcomes (M8, M9), adding PAVA to the predicted probabilities affects the selection of optimal dose. For method M10 which uses Gaussian process and maximizes the probability of utility higher than the fixed dose, the majority of the patients have optimal doses around the fixed dose. For method M11 which uses Gaussian process and maximizes the posterior mean of the utility function, the distribution of the optimal dose is most similar to the distribution from the true model.

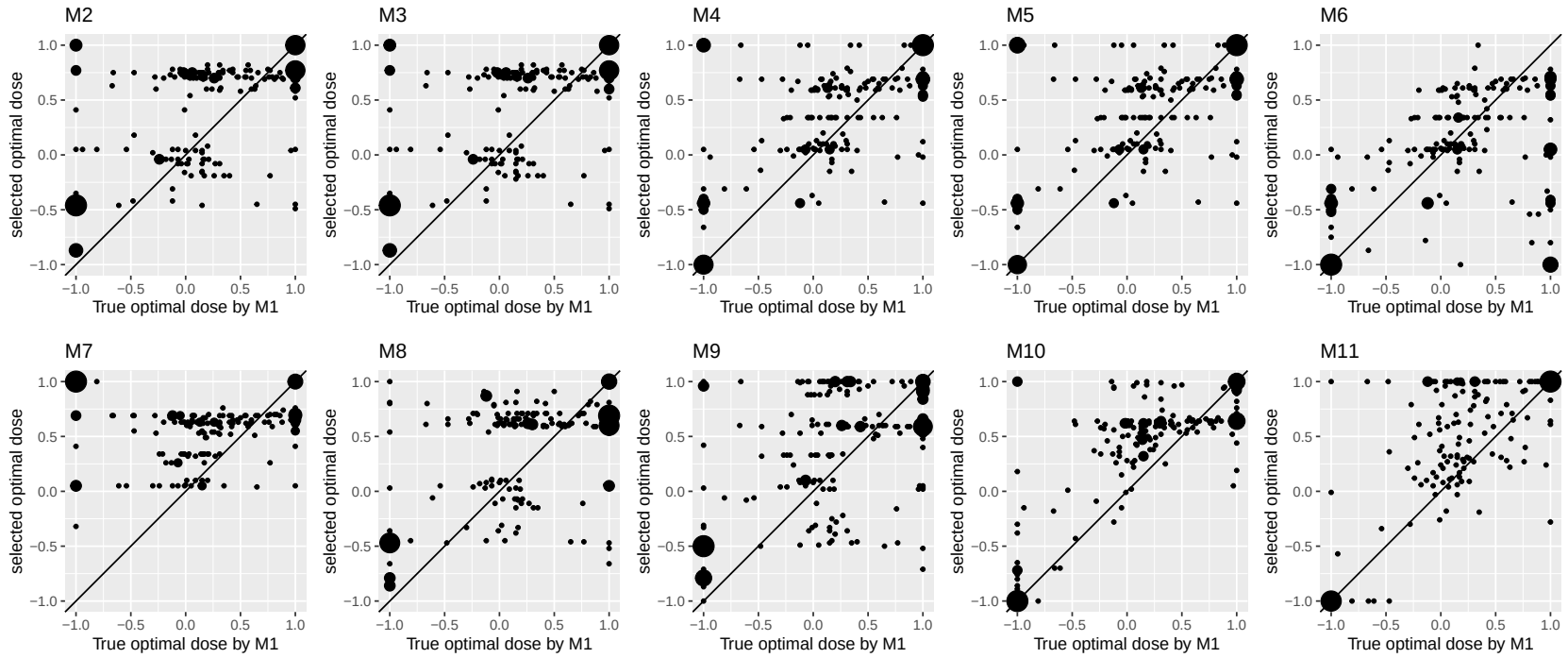

Figure 4.2: Distribution of optimal dose of n=200 patients for different methods under scenario 1 with utility 1

Table 4.6: Comparison of all methods under scenario 1 with utility 1,2,3

		True	RF		monotone RF				RF & PAVA		GP		FD
Method		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
U1	mean dose	0.221	0.206	0.219	0.233	0.245	-0.017	0.379	0.266	0.28	0.337	0.303	0.464
	sd dose	0.664	0.479	0.473	0.446	0.439	0.478	0.246	0.522	0.513	0.499	0.63	0
	mean E	0.703	0.623	0.625	0.628	0.629	0.532	0.646	0.649	0.651	0.674	0.687	0.658
	mean T	0.275	0.294	0.299	0.299	0.304	0.193	0.361	0.322	0.328	0.327	0.325	0.415
	mean $\bar{U}$	66.1	60.8	60.8	61	60.9	59.2	60	61.1	61	62.3	63.1	58.7
	% IP $\bar{U}$	1	0.278	0.27	0.3	0.286	0.053	0.17	0.32	0.308	0.484	0.595	0
U2	mean dose	0.017	0.027	0.027	0.03	0.029	-0.117	0.097	-0.009	-0.021	0.033	-0.001	0.142
	sd dose	0.667	0.495	0.495	0.469	0.469	0.483	0.255	0.561	0.539	0.476	0.662	0
	mean E	0.632	0.571	0.571	0.571	0.57	0.505	0.571	0.568	0.564	0.587	0.6	0.567
	mean T	0.191	0.223	0.222	0.218	0.218	0.161	0.253	0.213	0.211	0.222	0.208	0.301
	mean $\bar{U}$	72.1	67.4	67.4	67.6	67.6	67.2	65.9	67.7	67.6	68.3	69.6	63.3
	% IP $\bar{U}$	1	0.463	0.463	0.484	0.484	0.433	0.294	0.497	0.486	0.562	0.707	0
U3	mean dose	-0.162	-0.095	-0.084	-0.1	-0.086	-0.176	-0.072	-0.178	-0.174	-0.167	-0.215	-0.067
	sd dose	0.646	0.478	0.476	0.455	0.451	0.468	0.253	0.532	0.521	0.389	0.601	0
	mean E	0.556	0.526	0.529	0.524	0.526	0.489	0.516	0.499	0.499	0.508	0.516	0.5
	mean T	0.124	0.179	0.184	0.173	0.178	0.146	0.194	0.155	0.157	0.162	0.136	0.229
	mean $\bar{U}$	69.6	65.1	65	65.4	65.3	65.5	64.1	65.2	65.2	65.3	67	61.6
	% IP $\bar{U}$	1	0.432	0.419	0.469	0.452	0.484	0.313	0.445	0.441	0.459	0.668	0

Table 4.6 shows the summary statistics of the optimal dose, Efficacy, Toxicity and expectation of utility under scenario 1 with different utility functions. To compare the various methods on a common scale, we calculate the percent of possible increase in expected utility relative to fixed dose, denoted by % IP $\bar{U}$. Thus by definition the fixed dose approach has 0 percent improvement and the true models have 100 percent improvement with other methods generally falling in between these values. When using utility 1, $\omega_1 < \omega_2$, so on average the optimal dose is higher and as a result, average efficacy and average toxicity are both higher. Under utility 3, $\omega_1 > \omega_2$, the average optimal dose is lower, and average Efficacy and average Toxicity are lower. Under this scenario, within each utility function, use of the true models results in the highest average expected utility, and the fixed dose results in the lowest average of expectation of utility. The Gaussian process methods (M10, M11) have the highest mean utility values with the other methods having generally similar performance varying somewhat across utility matrices.

In practice, all covariates (e.g. biomarker panel) may have no true association with the patient outcomes so we also consider a null scenario (S0) in which the E and T models have only d without covariates x in them. Thus the dose effects are the same across all patients. As shown in Figure 4.3, the dose-only model has nearly as high of utility values as are possible using the true model. All other modeling approaches which consider other covariates, generally have poorer utility values than fixed dosing in this setting.

The comparison of expectation of utility under all scenarios is shown in Figure 4.3 and figures in the appendix, Fig C1, Fig C2. Table 4.7 shows the comparison of percentage of expectation of utility improvement under all scenarios with utility 1. In scenario 2, when there are only dose main effects and no dose-covariate interactions, the main effects of the covariates still play a role in optimal dose selection, and the Gaussian process which maximizes the posterior mean of utility function (M11) is

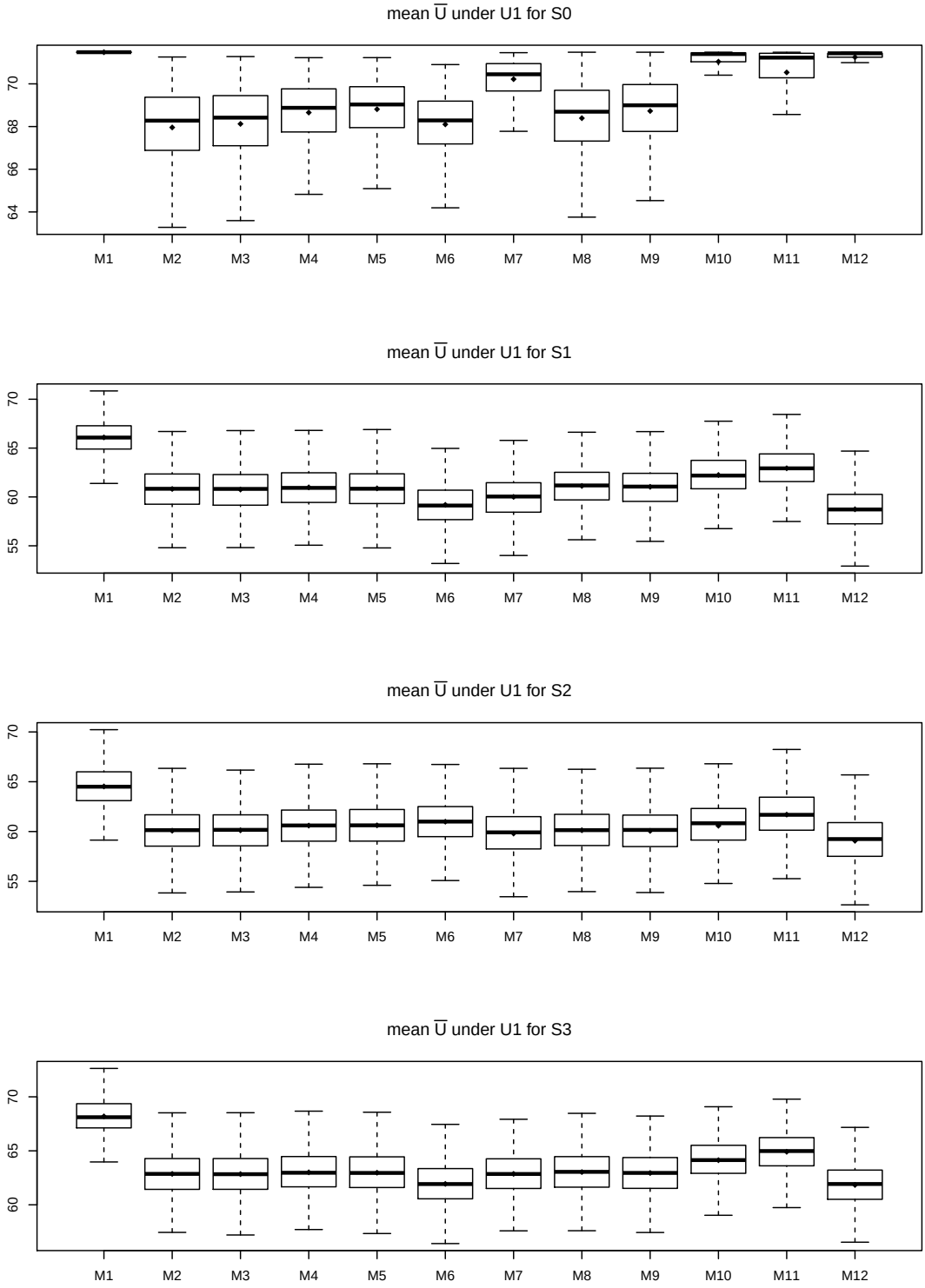

Figure 4.3: Simulation results for scenario 0, 1, 2, 3 under utility 1. Boxplot of population average of expectation of Utility for 1,000 simulation trials.

better than other methods. In scenario 3, when covariates are correlated, the Gaussian process methods (M11, M10) outperform the other methods. In scenario 4 and 5, with the increased number of noise covariates, the ARD feature of the Gaussian process and the horse-shoe priors help in model building and methods M12 and M13 have better performance than others, especially in scenario 5 when random forest (M2-M9) shows no improvement of utility function. In scenario 6, with the larger sample size, the magnitude of the improvement increases, and Gaussian process methods (M10, M11) still perform better than others. In scenario 7 when the covariates are all binary, to our surprise, the random forest methods (M2-M9) all have poor performance, while Gaussian process methods (M10, M11) still improve the utility function relative to fixed dosing (M12). In scenario 8, 9, 10, when the marginal models have covariate interactions or are mis-specified, all the methods have smaller magnitude of improvement, but the Gaussian process methods (M11, M10) still perform better than random forest methods (M2-M9), showing their robustness. Across all those scenarios, the copula makes no difference compared to assuming independence (M2 vs M3, M4 vs M5). Including the monotonicity in random forest leads to a small improvement in the mean utility function (M4 vs M2). However, adding monotonicity in categorical random forest afterwards does not help, as seen by comparing M9 vs M10. As shown in Figure 4.2, the optimal dose distribution of Gaussian process which maximizes posterior mean of utility function (M11) is most similar to the true model (M1), and also has the highest percent improvement in utility. Using Gaussian process modeling and selecting dose to maximize the posterior probability of higher utility than associated with fixed dosing (M10), tends to put a lot of patients near the fixed dose, but still improves the mean utility function compared to fixed dose.

Table 4.7: Comparison of percentage of utility function improvement under all scenarios with utility 1

	True	RF		monotone RF				RF & PAVA		GP		FD
Scenarios	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
S1	1	0.278	0.27	0.3	0.286	0.053	0.17	0.32	0.308	0.484	0.595	0
S2	1	0.155	0.162	0.247	0.251	0.324	0.123	0.163	0.151	0.247	0.471	0
S3	1	0.144	0.142	0.169	0.164	-0.003	0.152	0.175	0.161	0.35	0.464	0
S4	1	0.177	0.172	0.178	0.172	-0.019	0.097	0.22	0.213	0.451	0.536	0
S5	1	0.014	0.014	0.017	0.016	-0.114	0.008	-0.001	-0.004	0.308	0.29	0
S6	1	0.385	0.376	0.417	0.4	0.105	0.255	0.42	0.412	0.55	0.654	0
S7	1	-0.284	-0.28	-0.017	-0.028	-0.685	0.064	-0.405	-0.259	0.171	0.165	0
S8	1	0.215	0.212	0.242	0.235	0.101	0.155	0.223	0.212	0.411	0.525	0
S9	1	0.062	0.057	0.067	0.058	-0.547	0.103	0.185	0.194	0.268	0.437	0
S10	1	0.04	0.038	0.083	0.079	-0.429	0.134	0.102	0.107	0.259	0.384	0

4.5.3 Parametric vs non-parametric models

In the above simulation, we compare different non-parametric methods and their performance in optimal dose selection. In this section, we compare some of them with parametric models such as logistic regression (M13), LASSO (M14) and constrained LASSO (CLASSO, M15) proposed by Li.et al. All the parametric models are built on dose d, covariates x, and dose-covariate interactions dx for the marginals and we use copula to link them. The results are shown in Table 4.8. When the parametric models are correctly specified, they are more efficient and do better than the non-parametric models, as shown by S1, S4 and S7.

Furthermore, we evaluated more mis-specified scenarios when the complexity of the true model increases. The true efficacy and toxicity models can have more interactions such as covariate interactions xx and dose-covariate-covariate interactions dxx. The dose effect is $g(d)$ where $g(.)$ is a non-linear function. The efficacy model has $log(d+1)$ instead of d, so the efficacy increases faster at lower doses and tends to level off. Similarly, the toxicity model has $exp(d)$ instead of d, so the toxicity increases faster at high doses. In the limited mis-specified scenarios with detailed description in Table C1, the non-parametric models are comparable to and even outperform the parametric models as shown by S8, MS1-4.

The parametric models assume a finite set of parameters to estimate, the complexity of the model is bounded even if the amount of data is unbounded. On the contrary, the non-parametric models can be viewed as having an infinite dimension of parameters, and the amount of information that the parameters can capture about the data can grow as the amount of data grows, which is more flexible.

Table 4.8: Comparison of percentage of utility function improvement under selected scenarios with utility 1 for selected non-parametric and parametric methods

	True	RF			GP		FD	Parametric		
Scenarios	M1	M2	M4	M8	M10	M11	M12	M13	M14	M15
S1	1	0.278	0.3	0.32	0.484	0.595	0	0.759	0.804	0.846
S4: 15 noise x added to data	1	0.177	0.178	0.22	0.451	0.536	0	-0.140	0.679	0.705
S7: binary covariates x	1	-0.284	-0.017	-0.405	0.171	0.165	0	0.49	0.663	0.657
S8: The true models have x, xx, d, dx	1	0.215	0.242	0.223	0.411	0.525	0	0.627	0.642	0.689
MS1: The true models have x, xx, d, dx, dxx	1	0.063	0.110	0.078	0.071	0.241	0	0.210	0.253	0.281
MS2: The true models have $x, g(d), g(d)x$	1	0.302	0.327	0.282	0.227	0.532	0	0.279	0.331	0.417
MS3: The true models have $x, g(d), g(d)x, g(d)xx$	1	0.198	0.225	0.198	0.095	0.325	0	0.219	0.285	0.303
MS4: The true models have $x, xx, g(d), g(d)x, g(d)xx$	1	0.084	0.152	0.095	0.086	0.257	0	0.016	0.107	0.114

4.6 Application

In this section, we applied the proposed methods to a dataset of patients diagnosed with non-small cell lung cancer and were treated with radiation treatment. Patients with follow-up less than 1 year were excluded, leaving 109 patients in the dataset to be analyzed. Our primary measure of efficacy is lack of disease progression (including local, regional and distant progression). Of the 109 patients, 50 patients were free from progression at 2 years, and were considered as $E = 1$. Our binary toxicity measure is the occurrence of grade 2 or greater lung toxicity (pneumonitis). Of the 109 patients, 32 experienced this toxicity. There were 20 patients who had both efficacy and toxicity observed. The baseline clinical features we considered were age, current smoker status, Karnofsky Performance Status(KPS), T stage and N stage of the cancer, as shown in the Table 4.9. Patients in this study received tumor doses ranging from 45 to 96 Gy, partially due to the clinical trial on which they were enrolled and varying preferences of clinicians, as well as patient factors such as cancer stage, tumor location and performance status. In Radiation Oncology, the dose to the tumor site (most relevant dose for predicting efficacy) is different from the dose received by the normal lung tissue (most relevant for predicting lung toxicity). The ratio of these two dose values can be assumed to be constant for each patient but varies between patients due to tumor volume and location within the lung. We apply our methods to estimate the optimal tumor dose and calculate the implied normal lung dose for each patient using this fixed ratio.

For the utility matrix, we chose $\omega_1 = 0.4$ and $\omega_2 = 0.7$, considering an outcome of both efficacy and toxicity ($E = 1, T = 1$) as more favorable than an outcome of neither toxicity nor efficacy ($E = 0, T = 0$). For the random forest on the marginal distributions, we include tumor dose and clinical features in the efficacy model, and lung dose and clinical features in the toxicity model. For the random forest on categorical outcomes and Gaussian process, we include tumor dose and the ratio of lung

Table 4.9: Descriptive statistics of patients (n=109)

Variable	Mean	Range
Age (Years)	65.6	39.6 - 85.2
KPS	86	60 -100
Tumor dose (Gy)	71.6	45.7 - 96.1
Lung dose (Gy)	14.6	3.2 - 26.1
Variable	Category	Percentage
Smoking	Current	47
	Never or former	53
T stage	1	18
	2	23
	3	29
	4	30
N stage	0	22
	1	13
	2	44
	3	21

dose to tumor dose as well as clinical features. For the fixed dosing method, we build a multinomial logistic regression model on tumor dose and the ratio of lung dose to tumor dose.

The dose only model (M12) selects the highest dose as the best fixed dose (for all the patients), while other methods select individualized optimal dose for each patient, as shown in Fig 4.4. There is not much difference between the independence assumption or using the copula to link the two marginals, as shown by M2 vs M3 and M4 vs M5. Adding the individual level toxicity penalty (M6) lowers the recommended doses. Methods M7 with dose penalty tend to favor the estimated best fixed dose (M12) and recommend the highest dose for the majority of the patients. Use of the PAVA monotonicity algorithm on predictions from the random forest on the 4-category outcomes has some impact on the selected optimal dose (e.g. see medians for M8 vs M9). The Gaussian process methods recommend a wide range of doses to different patients, as shown by M10 and M11.

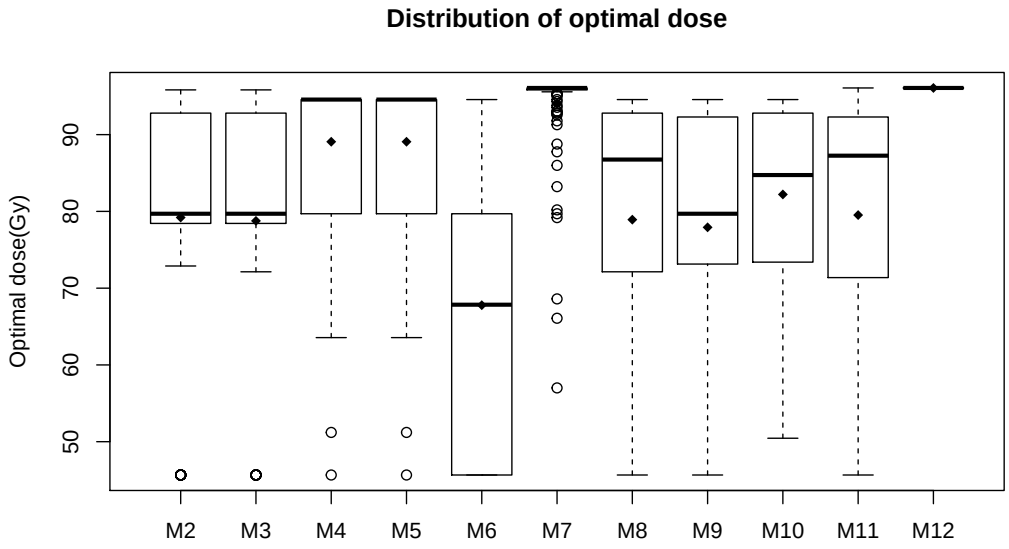

Figure 4.4: Boxplot of optimal doses by different methods for the 109 patients

To illustrate how the proposed utility calculations work we plot in Fig 4.5 the estimated probabilities of efficacy and toxicity and the corresponding utility functions versus dose for 3 selected patients. In the top row (a) we show the predictions from random forest models for efficacy and toxicity with copula to link them (M2), and the bottom plots (b) show the random forest models for efficacy and toxicity constrained to be monotone in dose and linked with copula (M4). Without forcing monotonicity, each of these 3 patients have regions of dose where as dose increases, the estimated probability of efficacy decreases, which is not plausible. Imposing monotonicity results in believable estimates and also smooths the dose-efficacy curves somewhat. The bottom panel (b) of the figure shows optimal doses for methods M4, M6, M7 which all utilize the same random forest models but differ in how they define optimal dose from these models. For patients 1 and 2, imposing monotonicity on the dose efficacy curve, results in a larger optimal dose. Patient 3 represents a patient with high baseline risk of toxicity exceeding what would typically be expected from dose only toxicity models. Because the probability of toxicity is higher than 0.3 even at the

lowest dose, method M6 with the individual toxicity penalty recommends the minimal dose. A higher dose is recommended from methods M4 and M7 however, as efficacy increases while toxicity plateaus. Plots showing other methods including M8, M9, M10, M11 on those three patients are included in the Appendix Fig. C3. The random forest for the categorical outcome results in non-monotone dose-efficacy and dose-toxicity relationship, and the PAVA which adjust the curves results in a plateau. The Gaussian process have smoother curves for dose-efficacy and dose-toxicity, with the limited sample size, the credible interval of the estimates is wide, as shown by the shadow around the posterior mean of utility.

The random forest method discretizes continuous variables and the curves are not as smooth as standard regression models, however in contrast to a parametric model, they can approximate any arbitrary shaped smooth function with sufficient data and number of splits.

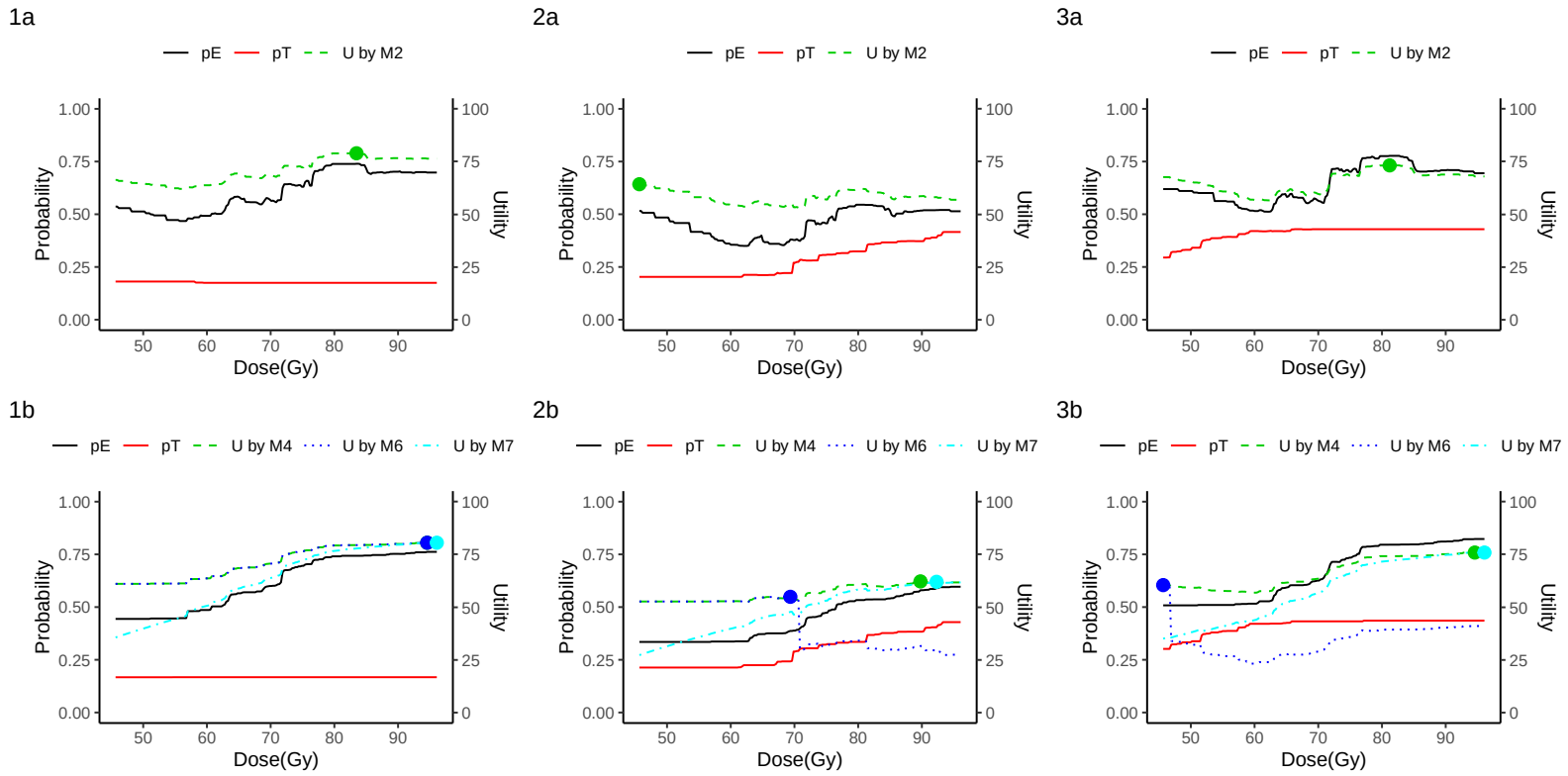

Figure 4.5: Optimal dose selected by different methods for three patients. Dose-efficacy and dose-toxicity curves are denoted by solid lines, expectation of utility values by different methods are denoted by dashed lines, optimal dose selected by different methods are denoted by points.

4.7 Discussion

In this paper, we propose to use flexible machine learning methods such as random forest and Gaussian process to build models for efficacy and toxicity depending on the dose and biomarkers. Copula is used to model the joint distribution of the two outcomes and the non-decreasing dose-efficacy and dose-toxicity relationship is constrained in the model building. A utility matrix with numerical utilities assigned to all patient outcome pairs allows the improvement in the utility due to a change in efficacy to depend on the level of toxicity. For each patient, the optimal dose is chosen to maximize the utility function or the posterior mean of the utility function. We further adjust the utility function with more constraints to incorporate clinical requirements, and consider the uncertainty in the estimation of the utility function by maximizing the posterior probability of utility function improvement.

In practice, the utility matrix can be pre-specified by clinicians, and evaluated by using simulations. The simulation may motivate the modification of utility values ω_1 and ω_2 in discussion with clinicians. Methods to obtain consensus utilities by summarizing questionnaires from a group of clinicians can also be used, such as the Delphi method [3]. Patient's preference and tolerance of the efficacy and toxicity outcomes can also be considered, which could help to build an individualized utility matrix. It is also possible to add more penalties to the utility function to take into consideration the financial cost of the treatment, which could depend on dose.

The expectation of the utility matrix $\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega})$ has been used to select the optimal dose for each patient. Given the observed data $\mathcal{D}_N$, frequentist estimation of $\hat{\boldsymbol{p}}(d,x)$ can be plugged into the utility function for optimization. Alternatively, Bayesian estimation will give the posterior distribution of $\boldsymbol{p}(d,x)$. The posterior mean of $\bar{U}(\boldsymbol{p}(d,x),\boldsymbol{\omega})$, which is the average of utility function over the posterior distribution, is used as the common Bayesian approach for optimal dose selection. However, the posterior mean of the utility function ignores the variance of the posterior distribution,

which reflects the uncertainty of the estimation. In this paper, we also propose to maximize the posterior probability that the utility function is higher than the utility function at a fixed dose. It is possible that the posterior mean of the utility function increases slightly with dose, but also has a larger variance, so there is not much improvement of the posterior probability of an improvement in the utility function. The posterior probability represents how confident we are about the improvement of the utility function, which can help clinicians in decision making. In the simulations, it is shown that using the posterior probability can help to avoid assigning extreme doses to patients, and shrink the individual optimal dose towards the population fixed dose.

Non-parametric methods such as random forest and Gaussian process provide flexible ways to model the outcome with all types of covariates. Some other tree based methods could also considered, such as BART [7] and gradient boosting trees [6]. Using Bayesian estimation, Gaussian process and BART are smoother and more accurate in prediction in terms of AUC and RMSE [25] compared to random forest and gradient boosting. In addition with the posterior distribution, they can provide uncertainty of the prediction and 95% credible intervals of estimation. The nature of the tree structure in random forest, BART and gradient boosting trees means these methods should be well suited to handle the interaction of dose and covariates. On the other hand, Gaussian process uses a Kernel function to describe the relationship across dose and covariates and should work better with continuous variables such as dose. To accommodate the non-decreasing dose-efficacy and dose-toxicity relationship, monotonicity is needed with respect to dose for all the values of x, that is for all patients, in the model. Monotonicity on a single covariate can be imposed by using virtual points in the Gaussian process approach [39], or as a splitting criteria in random forest and gradient boosting trees. But for BART, in current versions of the method, the monotonicity will be applied on all covariates [8], which is not

appropriate for this situation.

In general how to choose prior distributions is a big issue in Bayesian estimation. For the Gaussian process method, especially with a large number of covariates, we need to select an appropriate prior for ρ, the weight for each covariate in the kernel function. A Laplace prior, which is a Bayesian version of robust Lasso regression, did not perform well in our simulation study when we had a large number of covariates and sparsity. The spike-and-slab prior is appealing as it puts a substantial point-mass on zero, but can be computationally intensive with a large number of covariates. The horseshoe prior has been shown to have comparable performance to the spike-and-slab prior and is more computationally efficient because of its global parameter and local scales [4]. We adopted the horseshoe prior in our simulations and found it had good performance, even with a low number of covariates.

Nonparametric machine learning methods are flexible and they only make weak assumptions about the underlying functional form. But they generally require a large sample size to estimate the functions and may suffer from overfitting. Thus the sample size as well as the number of covariates and the expected complexity of the function should play a role in determining which method is appropriate in any setting. With smaller sample size, parametric machine learning methods such as neural networks can be considered as a powerful tool, or classical methods such as logistic regression could be appropriate.

The outcomes for defining the utility may be more complex than the binary ones for E and T considered in this paper. Utility values can be assigned to ordinal categorical efficacy and toxicity outcomes and the 2x2 utility matrix can be expanded. Survival outcomes can also be broken down into time intervals and these used to assess the trade-offs between quality and quantity of life, as in quality adjusted survival analysis [14].

CHAPTER V

Discussion

The goal in personalized medicine is to select individualized optimal treatment to improve the health outcome for each patient based on their characteristics such as biomarkers. In this dissertation, we consider the statistical challenges in personalized medicine and propose new methods to build and evaluate models for individualized treatment in oncology. The questions we are trying to answer include what is the structure of the model and which input feature or biomarker to include in the model. When the model is built, we need to evaluate how good the model is and whether it gives good predictions. How these predictions can be used to decide a treatment is the final step for individualized optimal dose finding.

In Chapter II, we demonstrate how to evaluate an existing model on a dataset with missing covariates. MI and IPW can be used to get unbiased BS and AUC if the imputation model or weight model is correctly specified. AIPW can improve the efficiency of IPW, and is double robust from mis-specification of the weight model or the imputing model. In the argument whether to include the outcome variable Y in models that handle the missing covariate, we find out Y should be included in the multiple imputation even if missingness does not depend on Y. On the other hand, if IPW or AIPW methods are used and missingness does not depend on Y, then it does not appear to be necessary to include Y in either the weight model or the missing

variable model. The approaches to handle missing data can result in fairly large variation in model performance estimates, as demonstrated by the prostate cancer data.

For the individualized treatment, we consider the setting with binary efficacy and toxicity outcomes, and the treatment consists of giving the dose of a therapy, where the dose can be any continuous value within a pre-specified range. The next two chapters focus on how to find the optimal dose for each patient using clinical features and biomarkers.

In Chapter III, we propose the constrained LASSO method to model the binary efficacy and toxicity outcomes using logistic link with dose, biomarkers and dose-biomarker interactions. The l_1 penalty of LASSO can help to select important biomarkers from the large number of potential clinical features. The constraint of the derivative of the model with respect to dose to be non-negative is used to maintain the non-decreasing dose-efficacy and dose-toxicity relationship. To maximize the rate of efficacy while limiting the rate of toxicity, we propose an optimal individualized dose finding rule by maximizing utility functions for individual patients, which is defined as a weighted combination of efficacy and toxicity probabilities. This approach can improve efficacy without increasing toxicity compared to the one dose for all approach.

In Chapter IV, we relax the assumption of independence of efficacy and toxicity in model building and use copula to model their joint distribution. The correlation of efficacy and toxicity might change the expectation of the utility, but has not much impact on the individualized optimal dose selection. Instead of parametric models as in Chapter III, we use flexible machine learning methods, which is becoming a trend in statistical analysis. The non-parametric approaches such as random forest and kernel machines such as Gaussian process can avoid model misspecification and be more robust. Monotonicity features can also be implemented in random forest and

Gaussian process to constrain the dose effect to be non-decreasing. The comparison of these methods to parametric models in this study reveals the pros and cons of them. In the optimal dose selection, we adjust the utility function with more constraints to meet the clinical requirement. The individual toxicity penalty helps to avoid unfavorable toxicity outcomes, and the dose penalty helps to avoid extreme doses assigned to patients. With the Bayesian estimation of Gaussian process, we also incorporate the uncertainty of estimation in the dose selection, which help to improve the utility function and also avoid extreme doses.

In Chapter III and IV, the outcomes we consider are binary efficacy and toxicity outcomes. In the lung data analysis, no progression in 2 years is treated as efficacy $E = 1$. Some patients with short follow-up and no progression observed were excluded from analysis because of this transformation of survival outcome to binary outcome. One point of future consideration is to extend the current methods to survival outcome. Parametric models such as Weibull model, semi-parametric models such as Cox proportional hazards model and non-parametric models such as survival random forest can be used to build models for progression conditional on dose and covariates. Copula can be used to link the survival model to other models to implement the joint distribution of efficacy and toxicity. To combine efficacy and toxicity in individualized optimal dose selection, the predicted progression free survival probability at a pre-specified time point can be plugged into the utility function. Alternatively, quality-adjusted survival can be used as a measure that captures the trade-offs between length and quality of life [14].

The current study focus on the dose assignment at the beginning of the study using the patient's baseline information. In a lot of medical practice, multiple treatments will be given sequentially rather than a single treatment. Whether to repeat a treatment that has obtained a favorable response or to modify it if the current response is unfavorable is a decision that has to be made. The choice of the next treatment is

often guided by updated data on the patients disease status and covariates, which are highly related to the previous treatment. Dynamic Treatment Regimes (DTR) have been proposed in the sequentially multiple assignment randomized trials (SMARTs) [48]. How to extend the proposed methods to the multiple-stage DTR is worth further investigation.

APPENDICES

APPENDIX A

Appendices for Chapter II

A.1 Differences between optimizing the likelihood, the AUC and the Brier Score

Brier score measures the mean squared difference between the predicted probability and the actual outcome of an event across all subjects. The lower the Brier score is for a set of predictions, the better the predictions are calibrated. When we evaluate an existing model such as a logistic model on the internal dataset, the Brier score will be minimized when the external model is the same as internal model, i.e, $F_E(Y|X) = F_I(Y|X)$.

Proof. Assume the $F_I(Y|X)$ as $expit(\alpha X)$ and $F_E(Y|X)$ as $expit(\beta X)$.

Brier score

$= \sum_Y \int_x (Y - \hat{p})^2 F_I(Y|X) F_I(X) dX$

$= \sum_Y \int_X (Y - \frac{1}{1+exp(-\beta X)})^2 (\frac{1}{1+exp(-\alpha X)})^Y (\frac{exp(-\alpha X)}{1+exp(-\alpha X)})^{(1-Y)} F_I(X) dX$

$= \int_X \left[(\frac{exp(-\beta X)}{1+exp(-\beta X)})^2 \frac{1}{1+exp(-\alpha X)} + (\frac{1}{1+exp(-\beta X)})^2 \frac{exp(-\alpha X)}{1+exp(-\alpha X)} \right] F_I(X) dX$

$= \int_X \frac{exp(-\alpha X)+exp(-\beta X)^2}{(1+exp(-\beta X))^2(1+exp(-\alpha X))} F_I(X)dX$

If for any X, $\frac{exp(-\alpha X)+exp(-\beta X)^2}{(1+exp(-\beta X))^2(1+exp(-\alpha X))}$ is minimized, then the integral over X will be minimized.

let $A = exp(-\alpha X)$, $B = exp(-\beta X)$, then the function can be written as

$$\frac{A + B^2}{(1 + B)^2(1 + A)}$$

Take derivative w.r.t B, we get:

$$\frac{2B(1 + B)^2(1 + A) - (A + B^2)2(1 + B)(1 + A)}{(1 + B)^4(1 + A)^2} = \frac{2(B - A)}{(1 + B)^3(1 + A)}$$

When $B < A$, the function will decrease, When $B > A$, the function will increase. Thus it will be minimized at $B = A$, i.e, when $F_E(Y|X) = F_I(Y|X)$.

□

AUC, which measures the area under the ROC Curve, indicates how well the predicted probabilities for the cases are separated from the controls. The question is under logistic models will the AUC be maximized when the external model is same as the internal model, i.e. $F_E(Y|X) = F_I(Y|X)$? The answer is it depends. The coefficients in the logistic regression model are not chosen to maximize the AUC, rather the coefficients are chosen to maximize the likelihood. In practice, these two sets of coefficients will frequently, but not always, be quite similar. However, if complete discrimination is possible, the maximum likelihood logistic regression coefficients will estimate the coefficients which separate the population [10, 37].

A.2 Consistency of IPW and AIPW estimators for Brier score

Considering the Brier score using the IPW method. Let

$$U_i(\theta, \gamma_1) = \theta R_i W_i - (Y_i - \hat{p_i})^2 R_i W_i,$$

where W_i depend on weight model with parameters γ_1.

Let $U_N(\theta, \gamma_1) = N^{-1} \sum_{i=1}^{N} U_i(\theta, \gamma_1)$, and it is straight forward that BS_{IPW} is the solution of $U_N(\theta, \gamma_1) = 0$. Let $U_E = E(U_N) = E(U_i(\theta, \gamma_1))$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{obs}, Y; \gamma_1)$. When the weight model is correctly specified, $\Pr(R = 1|X_{obs}, Y; \gamma_1^*) = \Pr(R = 1|X_{obs}, Y)$, then $E(R_i W_i) = 1$, and it is clear that $U_E(\theta, \gamma_1) = 0$. Because $U_N(\theta, \gamma_1)$ converges uniformly to $U_E(\theta, \gamma_1)$, BS_{IPW} is a consistent estimator.

The proof is similar for AIPW estimator. We first demonstrate consistency for a slightly modified estimator, which we call BS_{AIPW*} with

$$BS_{AIPW*} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{p_i})^2 R_i W_i + E[(Y_i - \hat{p_i})^2](1 - R_i W_i)$$

Let

$$V_i(\theta, \gamma_1, \gamma_2) = \theta - \left\{ (Y_i - \hat{p_i})^2 R_i W_i + E[(Y_i - \hat{p_i})^2](1 - R_i W_i) \right\},$$

where W_i depend on weight model with parameters γ_1 and $E[(Y_i - \hat{p_i})^2]$ depend on the model for missing covariates with parameters γ_2.

Let $V_N(\theta, \gamma_1, \gamma_2) = N^{-1} \sum_{i=1}^{N} V_i(\theta, \gamma_1, \gamma_2)$, then it is straightforward to see that BS_{AIPW*} is the solution of $V_N(\theta, \gamma_1, \gamma_2) = 0$. Let $V_E = E(V_N) = E(V_i(\theta, \gamma_1, \gamma_2))$. It is easy to see that $V_N(\theta, \gamma_1, \gamma_2)$ converges uniformly to $V_E(\theta, \gamma_1, \gamma_2)$, thus the solution

to $V_N(\theta, \gamma_1, \gamma_2) = 0$ converges to the solution of $V_E(\theta, \gamma_1, \gamma_2) = 0$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{obs}, Y; \gamma_1)$. When the weight model is correctly specified, $\Pr(R = 1|X_{obs}, Y; \gamma_1^*) = \Pr(R = 1|X_{obs}, Y)$, then $E(R_i W_i) = 1$.

Let γ_2^* be the probability limits of γ_2 using the model for the missing covariates $F(X_{mis}|X_{obs}, Y; \gamma_2)$. When the model is correctly specified, i.e., $F(X_{mis}|X_{obs}, Y; \gamma_2^*) = F(X_{mis}|X_{obs}, Y)$, then $E\{E[(Y_i - \hat{p}_i)^2] - (Y_i - \hat{p}_i)^2\} = 0$.

When either working model is correctly specified, it is clear that $V_E(\theta, \gamma_1, \gamma_2) = 0$, and that the θ that solves $V_E(\theta, \gamma_1, \gamma_2) = 0$ is the true BS. Because V_N converges uniformly to V_E, BS_{AIPW*} is a consistent estimator.

For the actual estimator BS_{AIPW} described in section 2.4 instead of calculating $E[(Y_i - \hat{p}_i)^2]$ where the expectation is over the distribution $F(X_{mis}|X_{obs}, Y; \gamma_2^*)$, we propose to use $(Y_i - \hat{p_i}^*)^2$ as an approximation.

A.3 Additional simulation results for correlated covariates

A.4 Implementing AIPW and IPW estimators when more than one variable has missing values

We propose the IPW and AIPW estimates of AUC and BS for single covariate missing in the main text and extend it here to more than one variable with missingness. We discuss how to build weight models and models for the missing covariates under different missing patterns.

First, we consider the block missing of covariates. Without loss of generality, consider the model with outcome Y and covariates X_1, X_2, X_3, and both X_2, X_3 are missing in some subjects. Let R_2 indicate X_2 is observed and R_3 indicate X_3 is observed, then $\Pr(R = 1) = \Pr(R_2 = 1, R_3 = 1)$. The weight model can be built by $\Pr(R = 1|X_1, Y)$ or $\Pr(R = 1|X_1)$, using the fully observed covariates with the

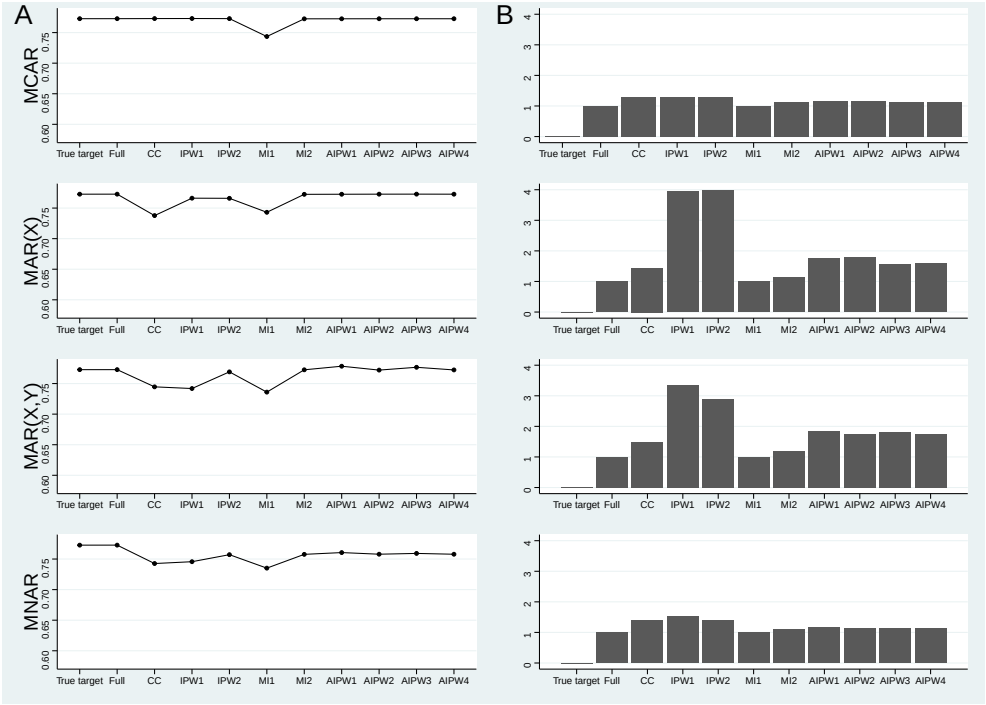

Figure A1: Simulation results of mean and relative SD of AUC for existing model M_1: $cor(X_1, X_3) = -0.5$. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

outcome or not. The models to impute X_2^* and X_3^* can be built separately, with $F(X_2|X_1, Y)$, $F(X_3|X_1, Y)$ or $F(X_2|X_1)$,$F(X_3|X_1)$ from the data of subjects with $R = 1$, and then obtain the predictions of X_2^* and X_3^* for all the subjects.

Next we look at a scattered pattern of missingness in the covariates. Use the same notation above with X_1 fully observed and X_2, X_3 are missing in some subjects. The weight model can be built by $\Pr(R = 1|X_1, Y)$ which indicate the complete cases without any missing, but may not capture the missingness for each covariate. Alternatively we can assume that the missingness of X_2 and X_3 are independent, then $\Pr(R = 1) = \Pr(R_2 = 1)\Pr(R_3 = 1)$. The weight models for R_2 and R_3 can be built separately by $\Pr(R_2 = 1|X_1, Y)$, $\Pr(R_3 = 1|X_1, Y)$ or $\Pr(R_2 = 1|X_1)$, $\Pr(R_3 = 1|X_1)$, using the fully observed covariates with the outcome or not. The models to impute X_2^* and X_3^* can be built separately as in block missingness.

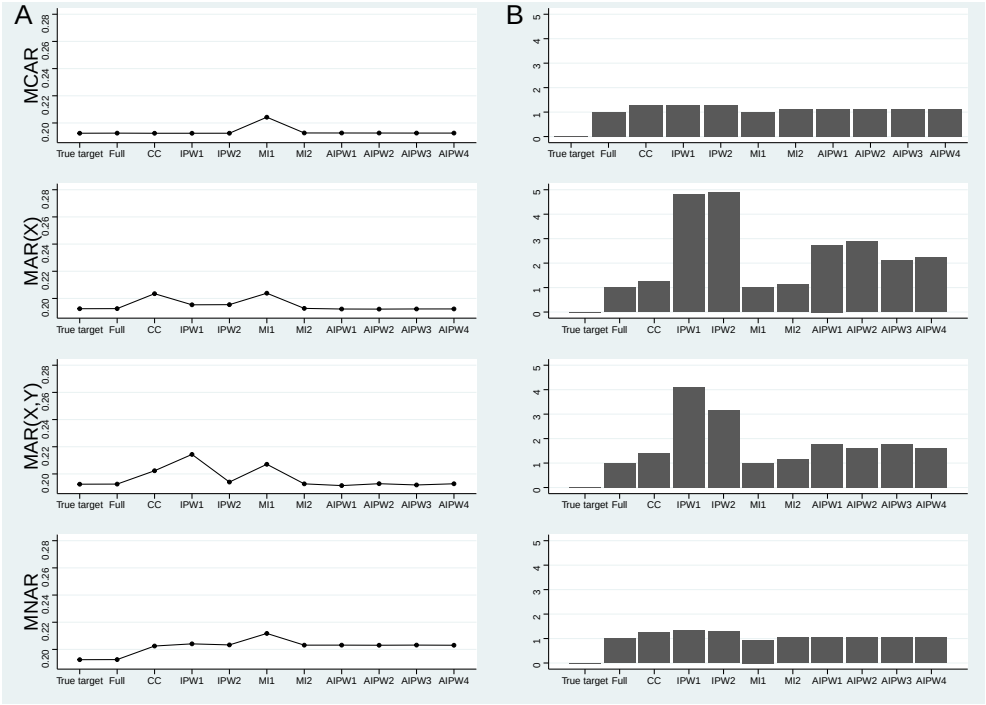

Figure A2: Simulation results of mean and relative SD of BS for existing model M_1: $cor(X_1, X_3) = -0.5$. Column A denotes mean BS. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

For the monotone missingness, X_1 is fully observed and both X_2, X_3 are missing in some subjects. For those with X_2 observed, X_3 is missing in some subjects too, with the probability of missing X_3 can depend on the value of X_2 under the MAR scenario. Now $\Pr(R = 1) = \Pr(R_3 = 1|R_2 = 1)\Pr(R_2 = 1)$ and we can build the model for R_2 using all the subjects and the model for R_3 using the subjects with X_2 observed. The models to impute X_2^* and X_3^* can be built separately as in block missing using the fully observed covariate X_1 with the outcome or not. Alternatively, the model to impute X_2^* can be built with $F(X_2|X_1, Y)$ or $F(X_2|X_1)$ from subjects with $R_2 = 1$ and get the predictions of X_2^* for all the subjects. Then the model to impute X_3^* can be built with $F(X_3|X_1, X_2, Y)$ or $F(X_2|X_1, X_2)$ from subjects with $R_3 = 1$ and get the predictions of X_3^* using X_2^* as the predictor covariate for all the subjects.

A.5 Simulation results when more than one variable has missing values

We consider the same model with true coefficients in M_1 and the covariates are independent. For block missing, similar as the single covariate missing, we consider the MCAR: block missing of X_2, X_3 has probability of 0.4; MAR (X_1): block missing of X_2, X_3 depends on the value of fully observed covariate X_1; MAR (X_1, Y): block missing of X_2, X_3 depends on the value of X_1, Y; MNAR: block missing of X_2, X_3 depends on the value of X_2, X_3. Fig A3 shows the simulation results after 1000 iterations for AUC, and the results are similar to Fig 2.1 for the single covariate missing situation.

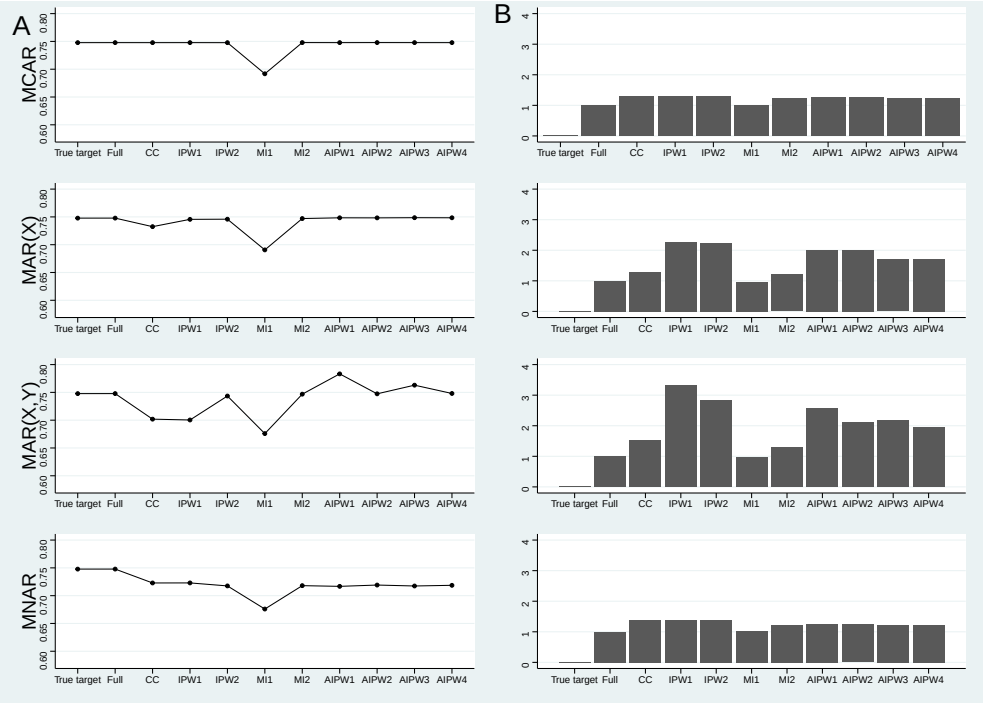

Figure A3: Simulation results of mean and relative SD of AUC for existing model with block missing of more covariates. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

For scattered missingness, we assume the missing of X_2 and X_3 are conditionally independent. For MCAR: missing of X_2 has probability of 0.4 and missing of

X_3 has probability of 0.2; MAR (X_1): missing of X_2 depends on the value of fully observed covariate X_1 and missing of X_3 depends on X_1 too with a different probability; MAR (X_1, Y): missing of X_2 and X_3 depends on the value of X_1, Y with different probabilities; MNAR: missing of X_2 depends on the value of X_2 and missing of X_3 depends on the value of X_3. As shown in Fig A4, under MCAR, MAR(X) and MAR(X,Y), the IPW and AIPW methods can get unbiased estimates when the models for $\Pr(R_2 = 1), \Pr(R_3 = 1)$ or the model to calculate X_2^*, X_3^* are correctly specified. But the variance are much higher in comparison to MI methods, especially for AIPW under MAR(X,Y).

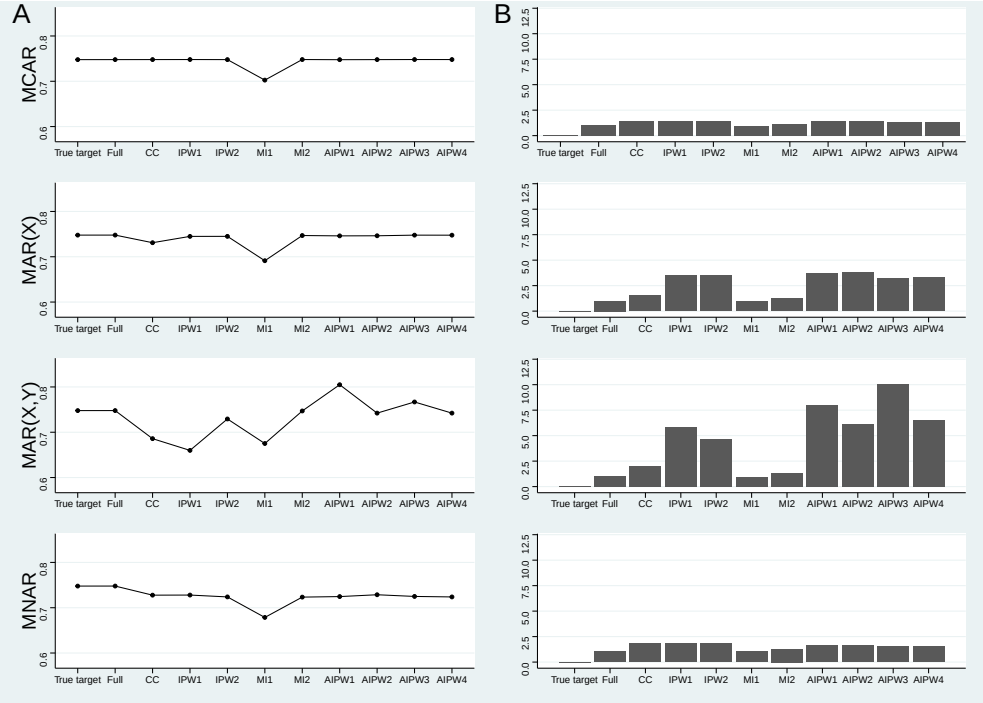

Figure A4: Simulation results of mean and relative SD of AUC for existing model with scatter missing of more covariates. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

For monotone missing, we assume the subjects with missing in X_2 have missing in X_3 and some subjects with X_2 observed have missing in X_3 too. For MCAR: missing of X_2 has probability of 0.4 and for those with X_2 observed, missing of

X_3 has probability of 0.5; MAR (X_1): missing of X_2 depends on the value of fully observed covariate X_1, and for those with X_2 observed, missing of X_3 depends on X_1 and X_2; MAR (X_1, Y): missing of X_2 depends on the value of X_1 and Y, and for those with X_2 observed, missing of X_3 depends on X_1, X_2 and Y; MNAR: missing of X_2 depends on the value of X_2, and for those with X_2 observed, missing of X_3 depends on the value of X_3. We compared different choices for the models to obtain X_3^*, either it includes X_2 or independent of X_2, and we saw no difference of the simulation results. In further simulations we saw that using X_2^* to predict X_3^* does not help when X_2, X_3 are correlated. As shown in Fig A5, under MCAR, MAR(X) and MAR(X,Y), the IPW and AIPW methods can get unbiased estimates when the weight model of $\Pr(R_2 = 1), \Pr(R_3 = 1|R_2 = 1)$ or the model to calculate X_2^*, X_3^* are correctly specified. The AIPW methods are more efficient than IPW methods.

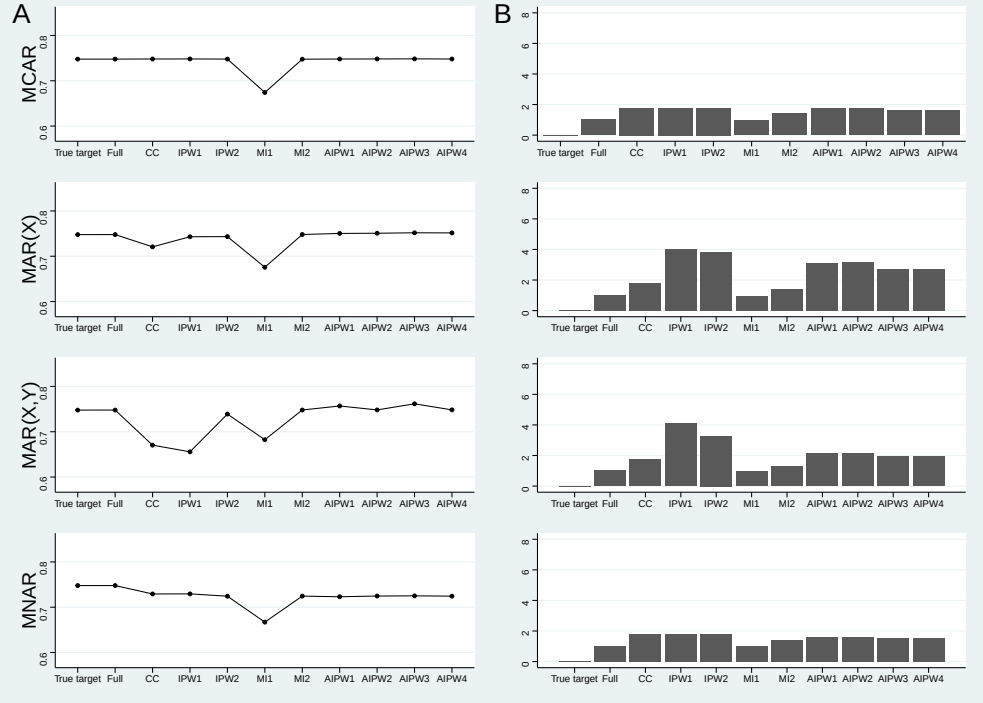

Figure A5: Simulation results of mean and relative SD of AUC for existing model with monotone missing of more covariates. Column A denotes mean AUC. Column B denotes SD relative to full data analysis. The four rows are different missingness mechanisms.

In conclusion, the extension of the IPW and AIPW methods to multiple covariates missing is feasible and have good performance under block missing and monotone missing.

APPENDIX B

Appendices for Chapter III

B.1 The minimizer to problem (3.4) always satisfies $\beta_j^+\beta_j^- = 0$ for $j = 1, ..., 2p+1$

Proof. Proof by contradiction: Consider the minimizer of (3.4) $\beta = \beta^+ - \beta^-$. Without loss of generality, assume we have $\beta_1^+ > 0, \beta_1^- > 0$. Consider another representation of the same $\beta = \tilde{\beta^+} - \tilde{\beta^-}$.

$\tilde{\beta_1^+} = \beta_1^+ - \min(\beta_1^+, \beta_1^-), \tilde{\beta_1^-} = \beta_1^- - \min(\beta_1^+, \beta_1^-)$

$\tilde{\beta_j^+} = \beta_j^+, \tilde{\beta_j^-} = \beta_j^-$ for $j = 2, ..., 2p+1$

Obviously, ($\tilde{\beta^+}$, $\tilde{\beta^-}$) satisfies the constraints of problem (3.4), and $\tilde{\beta_1^+} + \tilde{\beta_1^-} < \beta_1^+ + \beta_1^-, \tilde{\beta_1^+} - \tilde{\beta_1^-} = \beta_1^+ - \beta_1^-$.

Then the objective function (3.4) can be bounded as

$$
\begin{aligned}
&L(\beta^+, \beta^-) \\
&= -\sum_{i=1}^n \left\{ y_i(\beta_0 + X_i(\beta^+ - \beta^-)) - \log(1 + e^{\beta_0 + X_i(\beta^+ - \beta^-)}) \right\} + \lambda \sum_{j=1}^{2p+1} (\beta^+ + \beta^-) \\
&= -\sum_{i=1}^n \left\{ y_i(\beta_0 + X_i(\tilde{\beta^+} - \tilde{\beta^-})) - \log(1 + e^{\beta_0 + X_i(\tilde{\beta^+} - \tilde{\beta^-})}) \right\} + \lambda \sum_{j=1}^{2p+1} (\beta^+ + \beta^-) \\
&> -\sum_{i=1}^n \left\{ y_i(\beta_0 + X_i(\tilde{\beta^+} - \tilde{\beta^-})) - \log(1 + e^{\beta_0 + X_i(\tilde{\beta^+} - \tilde{\beta^-})}) \right\} + \lambda \sum_{j=1}^{2p+1} (\tilde{\beta^+} + \tilde{\beta^-}) \\
&= L(\tilde{\beta^+}, \tilde{\beta^-}).
\end{aligned}
$$

This contradicts with the assumption that (β^+, β^-) is the minimizer of (3.4). $\square$

B.2 Additional simulation results

Table B1: Simulation results for two toxicities. Summary of average Efficacy improvement compared with fixed dose with P(Toxicity1) constrained to be ≤ 0.2 and P(Toxicity2) constrained to be ≤ 0.23. Results from 1000 simulated trials. Each scenario true logistic models for E, T1 and T2 include main effect for the biomarkers, dose and biomarker-dose interactions, with coefficients as shown below.

Scenarios	Efficacy and Toxicity model coefficients												FS	LASSO	cLASSO	Possible Improvement	cLASSO > LASSO
		Biomarker					Dose	Interactions									
A0	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.581	0.549	0.606	0.121	66.3%
	T1	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					
A1	E	1	0	0	0	0	1	0	0	0	0	0	0.682	0.781	0.800	0.036	50.2%
	T1	-1	0	0	0	0	1	0	0	0	0	0					
	T2	-1	0	0	0	0	1	0	0	0	0	0					
A2	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.639	0.627	0.695	0.114	67.2%
	T1	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					
A3	E	1	0	0	0	0	1	.4	.4	.4	-.8	0	0.454	0.471	0.510	0.121	59.3%
	T1	-1	0	0	0	0	1	-.4	-.4	-.4	.8	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					
A0*	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.659	0.661	0.705	0.111	62.3%
	T1	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					

Table B1 Continued:

Scenarios	Efficacy and Toxicity model coefficients											FS	LASSO	cLASSO	Possible Improvement	cLASSO > LASSO
		Biomarker					Dose	Interactions								
A1*	E	1	.2	.3	.1	0	1	0	0	0	0	0	0.575	0.677	0.700	0.039	50.9%
	T1	-1	-.2	-.3	-.1	0	1	0	0	0	0	0					
	T2	-1	0	0	0	0	1	0	0	0	0	0					
A2*	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.702	0.667	0.731	0.125	55.1%
	T1	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					
A3*	E	1	.2	.3	.1	0	1	0	.2	-.1	.6	0	0.622	0.593	0.658	0.115	60.3 %
	T1	-1	-.2	-.3	-.1	0	1	0	-.2	.1	-.6	0					
	T2	-1	0	0	0	0	1	-.5	0	0	0	.5					

a. The intercept for Efficacy models is 0, for Toxicity1 models is -1.386, for Toxicity2 models is -1.2

b. Possible Improvement=$\mathbb{E}^{\mathcal{F}_{Theory}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)$, the percentage of improvement=$\{\mathbb{E}^{\mathcal{F}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}/\{\mathbb{E}^{\mathcal{F}_{Theory}}(E) - \mathbb{E}^{\mathcal{F}_{FD}}(E)\}$.

c. The random walk method of selecting θ was run for 1000 iterations.

d. Scenario A2, A2* have $cor(x_1, x_2, x_3) = 0.6$.

e. Scenarios A3, A3* have 15 noise covariates with coefficients 0 added.

APPENDIX C

Appendices for Chapter IV

C.1 Using copulas to link two marginal distributions

The marginal cumulative distribution function for a binary outcome Y, which could be E or T, is $F(0) = \Pr(Y \leq 0|d, x) = 1 - p, F(1) = \Pr(Y \leq 1|d, x) = 1$, where p is the probability of $Y = 1$ given d and x.

The copula approach makes use of the property that $(V_E, V_T) = (F(E), F(T))$ have uniformly distributed marginals. The copula is defined as

$$c(v_E, v_T) = \Pr(V_E \leq v_E, V_T \leq v_T) = \Pr(E \leq F^{-1}(v_E), T \leq F^{-1}(v_T))$$

So for the joint probability given d and x, $\Pr(E = 0, T = 0|d, x) = \Pr(F(E) \leq F(0), F(T) \leq F(0)) = \Pr(F(E) \leq 1 - p_E, F(T) \leq 1 - p_T)$, then

$$p_{00}(d, x) = \Pr(E \leq F^{-1}(1 - p_E), T \leq F^{-1}(1 - p_T)) = c(1 - p_E, 1 - p_T)$$

$$p_{10}(d, x) = 1 - p_T - c(1 - p_E, 1 - p_T)$$

$$p_{01}(d, x) = 1 - p_E - c(1 - p_E, 1 - p_T)$$

and

$$p_{11}(d,x) = c(1-p_E, 1-p_T) - p_E - p_T + 1.$$

Then for each subject, the outcome (E,T) is from the multinomial distribution with probabilities $p_{00}, p_{10}, p_{01}, p_{11}$ and the loglikelihood is

$$l = I(E=0,T=0)log(p_{00}) + I(E=1,T=0)log(p_{10}) + I(E=0,T=1)log(p_{01})$$

$$+I(E=1,T=1)log(p_{11}).$$

Different copula functions can be used. For this paper a Gaussian copula is adopted

$$c(v_E, v_T) = \Phi_2[\Phi^{-1}(v_E), \Phi^{-1}(v_T)|\alpha],$$

where α is the correlation parameter of $\Phi^{-1}(v_E)$ and $\Phi^{-1}(v_T)$ [43]. The marginal probabilities p_E, p_T can be specified by parametric models and parameters can be estimated jointly in the loglikelihood together with α. Two step estimation methods can also be used [21], where in the first step estimates are obtained for the parameters from the marginal models and are then held fixed in the second step, which consists of maximizing the likelihood to estimate the association function parameter. Joe and Xu (1996) showed that with the 2-step estimate for parameters is consistent and asymptotically normally distributed.

C.2 Additional results

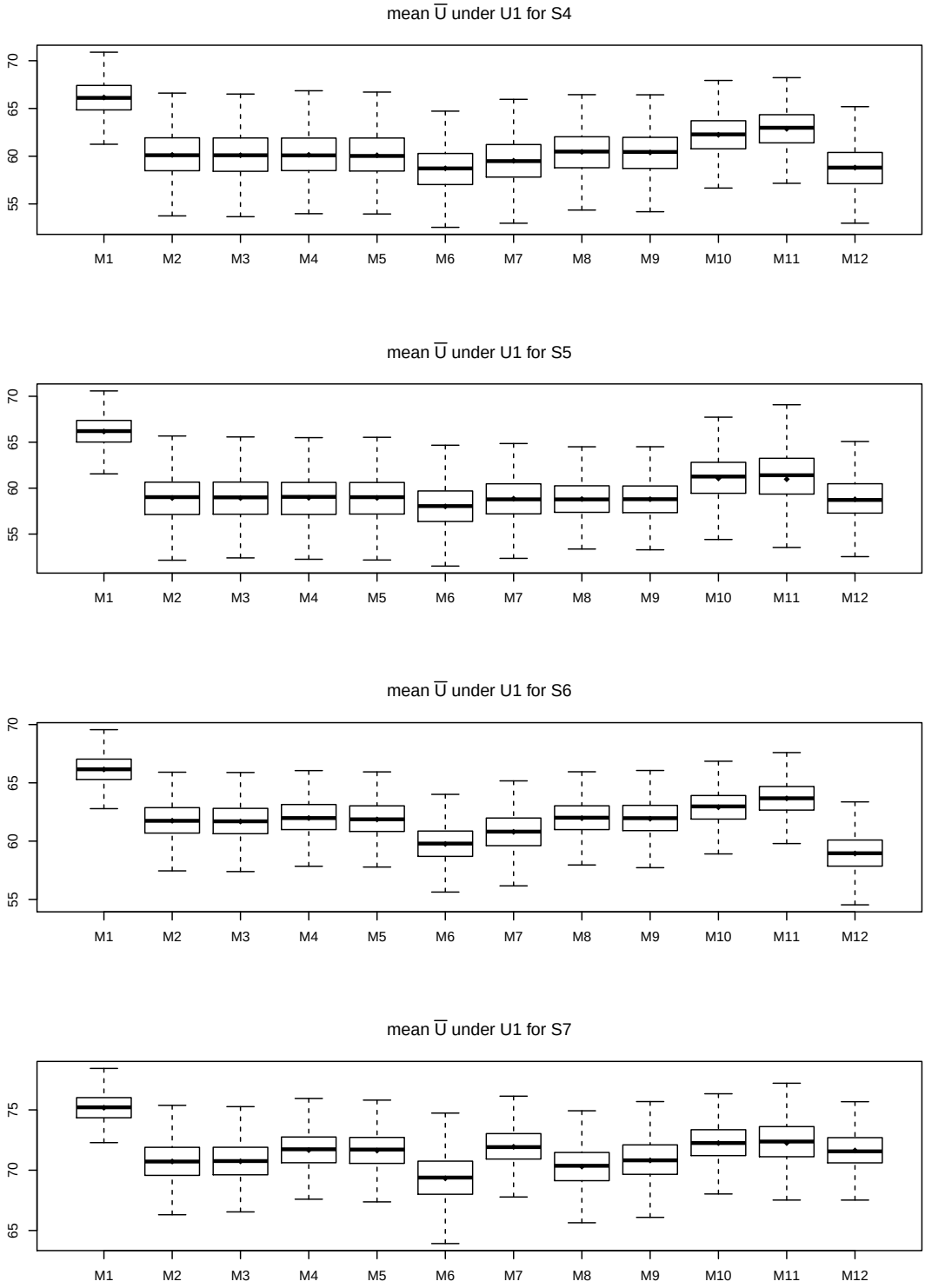

Figure C1: Simulation results for scenario 4-7 under utility 1. Boxplot of population average of expectation of Utility for 1,000 simulation trials.

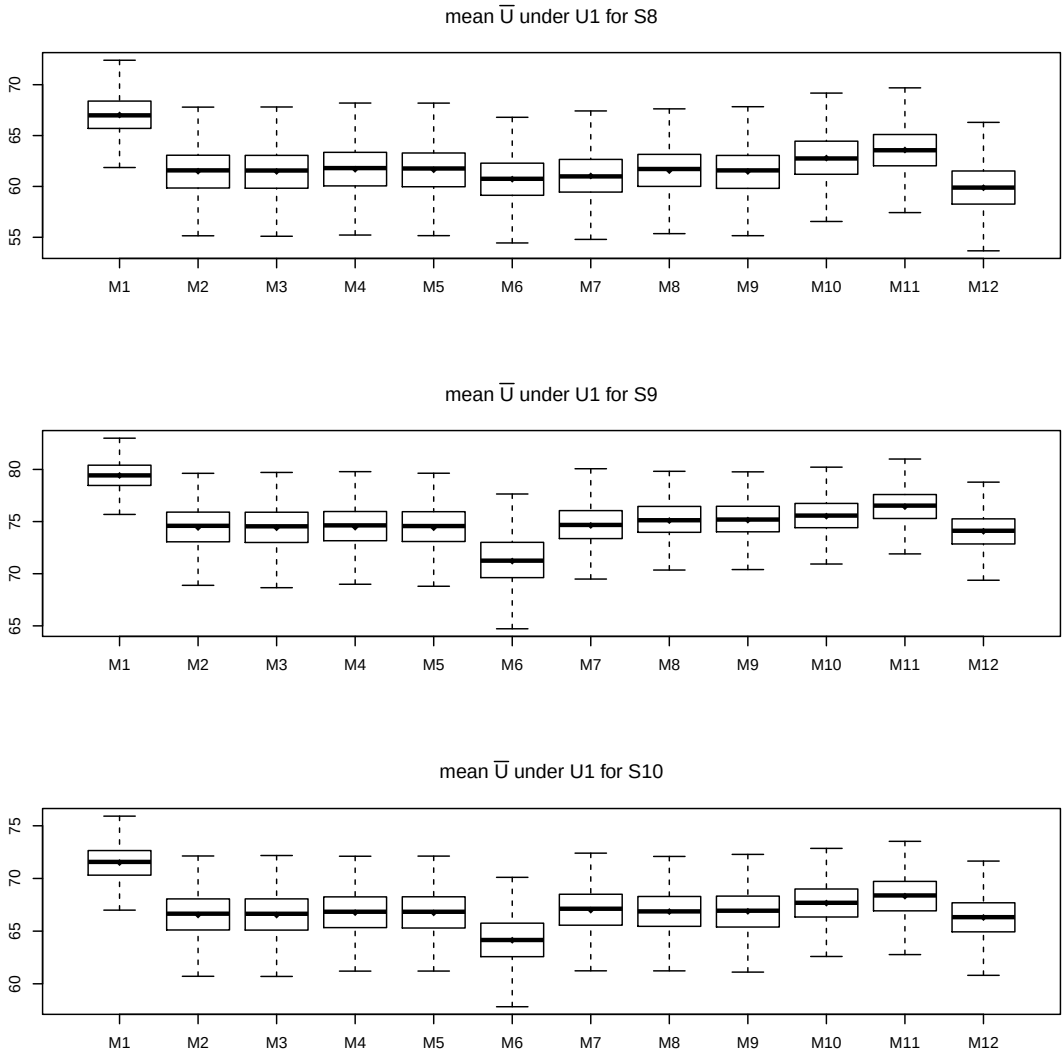

Figure C2: Simulation results for scenario 8-10 under utility 1. Boxplot of population average of expectation of Utility for 1,000 simulation trials.

Table C1: List of scenarios to compare parametric vs non-parametric models

S1	The true E & T models have x, d, dx
S4	S1 with 15 noise covariates $x_6, ..., x_{20}$ added to the data
S7	S1 with binary covariates x
S8	The true models have x, xx, d, dx Of the 5 covariates, coefficients are non-zero for 4 x, 1 xx, 4 dx
MS1	The true models have x, xx, d, dx, dxx Of the 5 covariates, coefficients are non-zero for 1 x, 1 xx, 1 dx, 2 dxx
MS2	The true models have $x, g(d), g(d)x$ Of the 10 covariates, coefficients are non-zero for 10 x, 1 $g(d)x$
MS3	The true models have $x, g(d), g(d)x, g(d)xx$ Of the 5 covariates, coefficients are non-zero for 1 x, 1 $g(d)x$, 2 $g(d)xx$
MS4	The true models have $x, xx, g(d), g(d)x, g(d)xx$ Of the 5 covariates, coefficients are non-zero for 1 x, 1 xx, 1 $g(d)x$, 2 $g(d)xx$

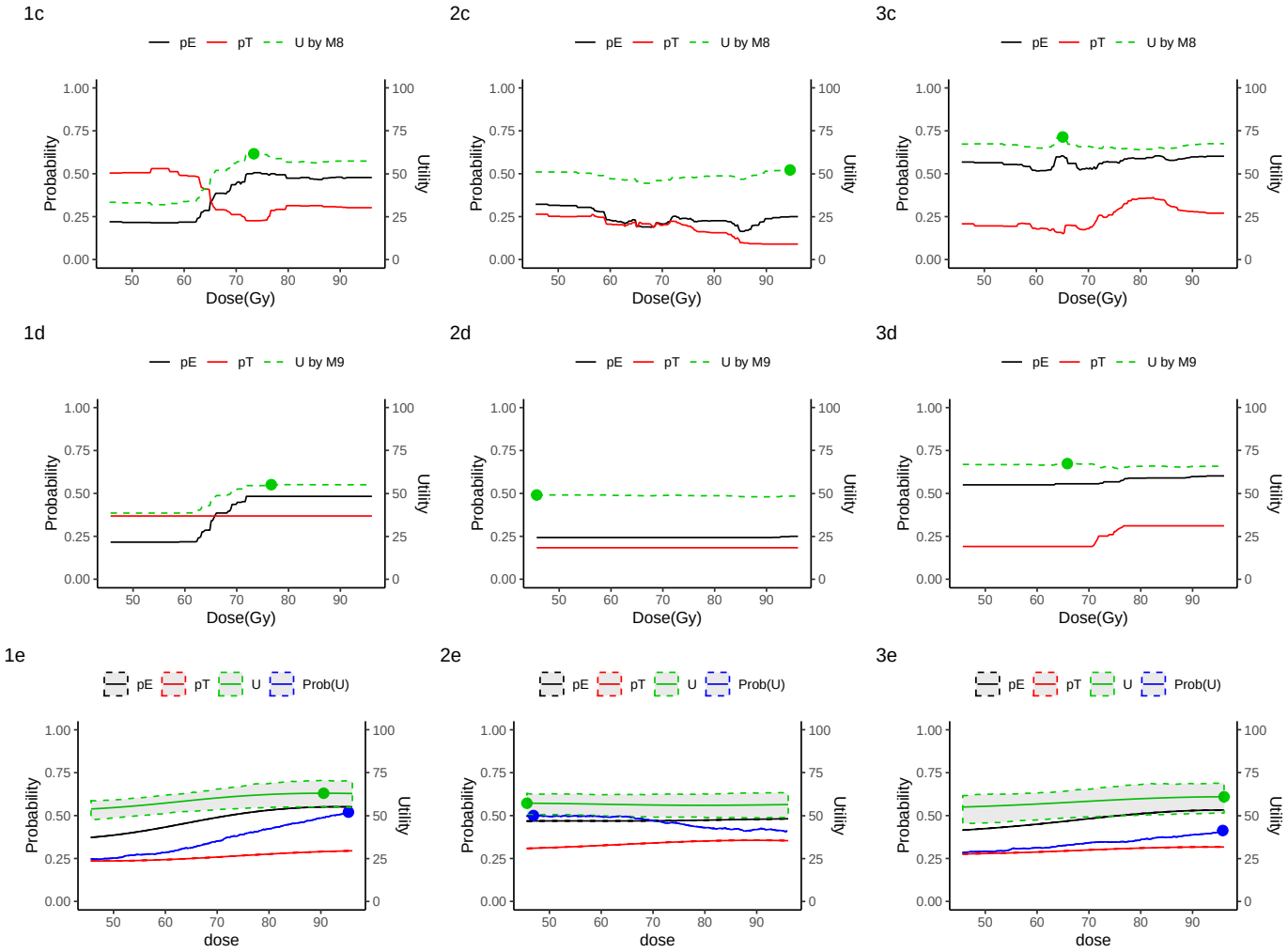

Figure C3: Optimal dose selected by different methods for three patients. Dose-efficacy and dose-toxicity curves are denoted by solid lines, expectation of utility values by different methods are denoted by dashed lines, optimal dose selected by different methods are denoted by points.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] Jonathan S Brajtbord, Michael S Leapman, and Matthew R Cooperberg. The capra score at 10 years: contemporary perspectives and analysis of supporting studies. *European urology*, 71(5):705–709, 2017.

[3] Robert H Brook, Mark R Chassin, Arlene Fink, David H Solomon, Jacqueline Kosecoff, and Rolla Edward Park. A method for the detailed assessment of the appropriateness of medical technologies. *International journal of technology assessment in health care*, 2(1):53–63, 1986.

[4] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009.

[5] Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[8] Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. High-dimensional nonparametric monotone function estimation using bart. *arXiv preprint arXiv:1612.01619*, 2016.

[9] Matthew R Cooperberg, David J Pasta, Eric P Elkin, Mark S Litwin, David M Latini, Janeen Du Chane, and Peter R Carroll. The university of california, san francisco cancer of the prostate risk assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. *The Journal of urology*, 173(6):1938–1942, 2005.

[10] Nicholas E Day and David F Kerridge. A general maximum likelihood discriminant. *Biometrics*, pages 313–323, 1967.

[11] FDA. Considerations for the design of early-phase clinical trials of cellular and gene therapy products. *Draft guidance for industry. Rockville, MD: Center for Biologics Evaluation and Research, FDA*, 2013.

[12] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[13] Brian R Gaines and Hua Zhou. Algorithms for fitting the constrained lasso. *arXiv preprint arXiv:1611.01511*, 2016.

[14] PP Glasziou, RJ Simes, and RD Gelber. Quality adjusted survival analysis. *Statistics in medicine*, 9(11):1259–1276, 1990.

[15] Beibei Guo and Ying Yuan. Bayesian phase i/ii biomarker-based dose finding for precision medicine with molecularly targeted agents. *Journal of the American Statistical Association*, 112(518):508–520, 2017.

[16] Tianhong He. Lasso and general l1-regularized regression under linear equality and inequality constraints. 2011.

[17] Leroy Hood and Stephen H Friend. Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nature reviews Clinical oncology*, 8(3):184, 2011.

[18] Jeroen Hoogland, Marit van Barreveld, Thomas PA Debray, Johannes B Reitsma, Tom E Verstraelen, Marcel GW Dijkgraaf, and Aeilko H Zwinderman. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*.

[19] Nadine Houede, Peter F Thall, Hoang Nguyen, Xavier Paoletti, and Andrew Kramar. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase i/ii trials. *Biometrics*, 66(2):532–540, 2010.

[20] Kristel JM Janssen, Yvonne Vergouwe, A Rogier T Donders, Frank E Harrell Jr, Qingxia Chen, Diederick E Grobbee, and Karel GM Moons. Dealing with missing predictor values when applying clinical prediction models. *Clinical chemistry*, 55(5):994–1001, 2009.

[21] Harry Joe and James Jianmeng Xu. The estimation method of inference functions for margins for multivariate models. 1996.

[22] Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

[23] Pin Li, Jeremy MG Taylor, Spring Kong, Shruti Jolly, and Matthew J Schipper. A utility approach to individualized optimal dose selection using biomarkers. *Biometrical Journal*, 62(2):386–397, 2020.

[24] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[25] Antonio R Linero. A review of tree-based bayesian methods. *Communications for Statistical Applications and Methods*, 24(6), 2017.

[26] Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[27] Roderick JA Little and Donald B Rubin. Statistical analysis with missing data. *Hoboken, NJ: Wiley*, 1987.

[28] Brent R Logan, Rodney Sparapani, Robert E McCulloch, and Purushottam W Laud. Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. *Statistical methods in medical research*, 28(4):1079–1093, 2019.

[29] Qi Long, Xiaoxi Zhang, and Brent A Johnson. Robust estimation of area under roc curve using auxiliary variables in the presence of missing biomarker values. *Biometrics*, 67(2):559–567, 2011.

[30] Junsheng Ma, Brian P Hobbs, and Francesco C Stingo. Statistical methods for establishing personalized treatment rules in oncology. *BioMed research international*, 2015, 2015.

[31] Lu Mao. On causal estimation using-statistics. *Biometrika*, 105(1):215–220, 2017.

[32] Guillermo Marshall, Bradley Warner, Samantha MaWhinney, and Karl Hammermeister. Prospective prediction in the presence of missing data. *Statistics in medicine*, 21(4):561–570, 2002.

[33] Karel GM Moons, Rogier ART Donders, Theo Stijnen, and Frank E Harrell Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101, 2006.

[34] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[35] John M Neuhaus and Nicholas P Jewell. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80(4):807–815, 1993.

[36] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[37] Margaret Sullivan Pepe and Mary Lou Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140, 2000.

[38] S Postel-Vinay, HT Arkenau, D Olmos, J Ang, J Barriuso, S Ashley, U Banerji, J De-Bono, I Judson, and S Kaye. Clinical benefit in phase-i trials of novel molecularly targeted agents: does dose matter? *British journal of cancer*, 100(9):1373, 2009.

[39] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 645–652, 2010.

[40] Matthew J Schipper, Jeremy MG Taylor, Randy TenHaken, Martha M Matuzak, Feng-Ming Kong, and Theodore S Lawrence. Personalized dose selection in radiation therapy using statistical models for toxicity and efficacy with dose and biomarkers as covariates. *Statistics in medicine*, 33(30):5330–5339, 2014.

[41] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

[42] M Seligman. Rborist: extensible, parallelizable implementation of the random forest algorithm. *R Package Version*, 1(3):1–15, 2015.

[43] Peter X-K Song, Mingyao Li, and Ying Yuan. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68, 2009.

[44] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[45] Stan Developent Team et al. Rstan: the r interface to stan. *R package version*, 2(1), 2016.

[46] Peter F Thall and John D Cook. Dose-finding based on efficacy–toxicity tradeoffs. *Biometrics*, 60(3):684–693, 2004.

[47] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[48] Lu Wang, Andrea Rotnitzky, Xihong Lin, Randall E Millikan, and Peter F Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107(498):493–508, 2012.

[49] Yuanjia Wang, Haoda Fu, and Donglin Zeng. Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, (just-accepted), 2017.

[50] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[51] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[52] Angela M Wood, Patrick Royston, and Ian R White. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632, 2015.

[53] Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

[54] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

[55] Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

[56] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.