# Quantifying the Utility of Personalized Treatment Decision Rules: Extending and Comparing Two Metrics for Summarizing the Heterogeneity of Treatment Effects

by

Yuan Xia

B.Sc., The University of British Columbia, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

**Master of Science**

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

April 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

**Quantifying the Utility of Personalized Treatment Decision Rules: Extending and Comparing Two Metrics for Summarizing the Heterogeneity of Treatment Effects**

submitted by **Yuan Xia** in partial fulfillment of the requirements for the degree of **Master of Science** in **Statistics**.

**Examining Committee:**

Paul Gustafson, Department of Statistics, UBC
*Supervisor*

Mohsen Sadatsafavi, Faculty of Pharmaceutical Sciences, UBC
*Supervisory Committee Member*

Lang Wu, Department of Statistics, UBC
*Supervisory Committee Member*

# Abstract

The treatment benefit prediction model is a type of clinical prediction model that quantifies the magnitude of treatment benefit given an individual's unique characteristics. As the topic of treatment effect modelling is relatively new, quantifying and summarizing the performance of treatment benefit models are not well studied. The 'concordance-statistic for benefit' and the 'concentration of benefit index' are two newly developed metrics that evaluate the discriminative ability of the treatment benefit prediction. However, the similarities and differences between these two metrics are not yet explored. We compare and contrast the metrics from conceptual, theoretical, and empirical perspectives and illustrate the application of the metrics. We consider the common scenario of a logistic regression model for a binary response developed based on data from a randomized controlled trial with two treatment arms. This thesis provides two major contributions: first, the two metrics are expanded into three pairs of metrics, each having a particular scope; second, it provides results of theoretical and simulation studies that compare and contrast the construct and empirical behaviour of these metrics. We found that the heterogeneity of treatment effect appropriately influences these metrics. Metrics related to the 'concordance-statistic for benefit' are sensitive to the unobservable correlation between counterfactual outcomes. In a case study, we quantify the metrics in a randomized controlled trial of acute myocardial infarction therapies on 30-day mortality. We conclude that these metrics help understand the heterogeneity of treatment effect and the consequent impact on treatment decision-making.

# Lay Summary

The treatment benefit prediction model helps identify which patients will benefit from treatment. The 'concordance-statistic for benefit' and the 'concentration of benefit index' are two similar metrics that measure the discriminative ability of the treatment benefit prediction. This thesis provides two major contributions: extended definitions for the two metrics and results of the comparisons. We conclude that the metrics help understand the heterogeneity of treatment effect and the consequent impact on treatment decision-making.

# Preface

This thesis presents original research conducted by Yuan Xia under the supervision of Professor Paul Gustafson and Professor Mohsen Sadatsafavi. All simulations and analyses were carried out by the author. This thesis was proofread by Professor Lang Wu in an official capacity, and the final version was approved.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgements

I am deeply and sincerely indebted to my supervisor, Professor Paul Gustafson, and my co-supervisor, Professor Mohsen Sadatsafavi, for their advice and guidance throughout my research. It was an honour to work with both of them, and this thesis could not have happened without their continuous support. I express sincere gratitude to my second reader, Professor Lang Wu, who has always been approachable throughout my graduate study and willing to help in many areas. Their help has benefited me both professionally and personally, and I am forever grateful.

Many thanks to the instructors and professors of the UBC statistics department for their excellent teaching and mentorship. I want to thank all my friends, fellow graduate students and staff members in the department for their help. Lastly, I want to thank my family for encouraging me and providing me with an abundance of kind support all of my life.

# Chapter 1

# Introduction

Clinical prediction models that objectively quantify the risk of an outcome or the benefit and harms of therapies are cornerstones of evidence-based medical practice. The prediction model is a statistical modelling method that uses patient characteristics to predict the expected value of a clinical outcome [3]. While the outcome can be a continuous variable or the rate of an event, the most common scenario is predicting the probability of a binary outcome (e.g., an event occurring or not) during a defined time horizon. Such models are conventionally referred to as risk prediction models. For example, these models can study the mechanical failure of artificial heart valves [4] and the incidence of cutaneous melanoma [5] based on patient characteristics.

After the prediction model is developed, it is essential to evaluate the validity of the model. The model's predictive performance is commonly quantified in terms of overall fit, calibration, and discrimination [6] [7]. The $R^2$ type metrics measure the overall fit by quantifying how close predictions are to the actual outcome. The calibration refers to the agreement between predictions and the observed outcomes values. It is summarized by metrics such as the calibration slope, the calibration-in-the-large, and the observed/expected risk ratio. The discrimination examines how well predictions discriminate between patients with low versus high risk of experiencing the outcome. The measurements such as the area under the curve (AUC), the concordance statistic (c statistic), Somer's D, and discrimination slope are widely used to quantify the discriminatory ability of risk prediction models [7].

In medical literature, treatment effect refers to the causal effect of a binary treat-

ment variable on an outcome of interest. Understanding heterogeneity of treatment effect (HTE) supports treatment-decision making via knowing whether the efficacy of treatment varies across different individuals [8]. Suppose individual variability in the direction and magnitude of treatment effects is well understood. In that case, treatment decision rules can accurately target patients who will benefit the most from the active treatment and avoid individuals with harmful treatment effects. For a new treatment compared with the current treatment, treatment decision rules can identify an appropriate study population for which the new treatment provides a significant benefit that compensates the cost [9]. Unfortunately, the treatment effect at the individual level can never be precisely measured because only one of the potential outcomes is observable in reality. The concept of HTE emphasizes the importance of predicting the expected treatment effect within identifiable subgroups. Subgroup analysis is a classical method of predicting HTE in a randomized controlled trial. In the presence of HTE, predicting the expected treatment effect for a given subgroup provides more clinical utility compared with predicting the risk [10]. This thesis is closely related to treatment effect modelling, which is a direct estimation of treatment benefit by including treatment assignment variables and potentially treatment-by-covariate interactions in a model. Thus, the treatment effect refers to treatment benefit.

There is a recent interest in directly modelling the treatment benefit given a patients' characteristics. However, metrics that are used to summarize the performance of a risk prediction model are not directly applicable to treatment benefit models. There is a need to formulate and evaluate metrics that communicate the performance of such treatment benefit models. Recently, two metrics are proposed for the discriminatory performance of these models, which are the concordance statistic for benefit $(cfb)$ and the concentration of benefit $(C_b)$. The $cfb$ and the $C_b$ are developed by Van Klaveren at el (2018) [11] and Sadatsafavi at el (2020) [12], respectively. The initial motivations of developing these two metrics are slightly different but end up reaching similar purposes. Thus, the primary goal of this thesis is to compare and contrast $cfb$ and $C_b$ from conceptual, theoretical, and empirical perspectives.

This thesis is laid out in the following manner: Chapter 1 provides the background. Chapter 2 provides a broad review of the risk modelling approaches and discriminative measures of risk models related to the two proposed metrics. Then, we turn to the model, which aims to predict the treatment benefit. This chapter also summarizes the initial $cfb$ proposed in [11] and $C_b$ in [12], and lays out the concerns of the currently deficient definitions of the metrics. Chapter 3 expands the theoretical definitions of $cfb$ and $C_b$ into three pairs of metrics: $(cfb, C_b)$, $(cfb_h, C_{b,h})$, and

$(\widehat{cfb_h}, \widehat{C}_{b,h})$, each devoted to a particular scope. The first pair of metrics summarize the genuine HTE as a function of covariates of interest in a target populaion, which is the final goal being earnestly pursued. The second pair of metrics seem more feasible compared with the first one. It captures the discriminatory performance of a previously-specified treatment benefit. The last pair of metrics contain the estimators of the second one. We discuss connections between the $cfb$ and the c statistic, and connections between $C_b$ and the Gini-like index [13]. The relationship between two population-level pairs of metrics is studied through stylized examples. Chapter 4 further explores the three pairs of metrics via simulation studies. In particular, the behaviour of three pairs of metrics is studied under different population HTE and various correlation levels of counterfactual outcomes and covariates. In Chapter 5, the metrics are applied to the acute myocardial infarction data from the GUSTO I randomized controlled trial [14]. The trial began in 1990 and was completed in 1993. Finally, Chapter 6 summarizes the thesis and touches on further possible directions for the research.

# Chapter 2

# Risk and treatment benefit prediction models

Chapter 1 provided an overview of risk models and treatment benefit models, along with their performance measurement. This chapter first takes a detailed look at the risk prediction model and related metrics that assess the model's discriminatory ability. There are connections between the $cfb$ and the c statistic and between the $C_b$ and the Gini index. Thus, we provide an overview of the c statistic and the Gini index before we direct our attention to the treatment benefit model and $cfb$ and $C_b$. These two metrics are proposed for measuring the predictive performance of treatment effect models.

## 2.1 Risk prediction models

A risk prediction model is used to predict the probability of a clinical event occurrence. We start by giving an example. In the first trimester of pregnancy, if a patient ends a pregnancy by medical abortion, what is the risk of incomplete abortion? For clinicians, it is essential to identify patients with high risks of incomplete abortion and then suggest a referral for further evaluation. Like this example, many medical decisions are binary. Hence, this thesis focuses primarily on binary response variables with two levels, either 'yes' or 'no.' We are often interested in predicting this binary clinical outcome as a function of prognostic variables (covariates). Before the modelling strategy is detailed, we first set up the notation. The binary response

is denoted as $Y$, with $Y = 1$ representing the occurrence of an event, and $Y = 0$ otherwise. Let $X$ denote a set of covariates.

Each covariate could be either quantitative or qualitative variables. Thus, it might take any value on the space of real numbers. However, the risk is a probability, bounded between 0 and 1. More specifically, the risk is a conditional probability of $Y = 1$ given known values of $X$. As is commonly the case, we consider a logistic regression analysis to model the relationship between a binary response and covariates, and a logit function is used to link the risk and the linear combination of covariates:

$$\text{logit}\left(E[Y|X]\right) = X\beta, \tag{2.1}$$

where $\beta$ is a column vector containing coefficients; $X$ represents a design matrix. As the response variable, $Y$, is Bernoulli distributed, the expectation of $Y$ equals the probability of the occurrence of $Y$.

Model parameters $\beta$ of a predictive model are estimated by the maximum likelihood method or a penalized maximum likelihood method given observed data $(Y, X)$. We denote the parameter estimates as $\hat{\beta}$. Once the model is determined, we can predict a new patient's risk according to their prognostic variables.

Returning to the medical abortion example, the common symptoms and side effects after medicine intake are possible prognostic variables. We assume that the log odds of abortion completion and the prognostic variables are linearly associated. A logistic regression model can then be used to model the association based on the historical data. This model can also predict a new patient's risk of incomplete abortion according to his/her related symptoms.

In the hopes of having accurate predictions, we would select a model based on its predictive performance. This performance is conventionally measured in three aspects: overall fit, discrimination, and calibration [6]. Take the example discussed in this section as an example. The calibration measures how accurate the predicted incomplete abortion risk is compared to the actual risk of incomplete abortion. The discrimination measures whether these predicted incomplete abortion risks correctly reflect the information about the occurrence of incomplete abortion. There are many techniques for assessing the predictive performance of a risk prediction model. The focus of this thesis is on the discriminatory performance of the models. Section 2.2 mainly focuses on the c statistic and the Gini index.

## 2.2 Performance metrics of risk prediction models

### 2.2.1 The receiver operating characteristic curve and the c statistic

For a binary response case, the outcome prediction can be interpreted as a classification process. This process combines the risk prediction and response classification. We classify a patient's response based on the predicted risk and a predetermined threshold $T$. For patient $i$, let $R_i$ be the risk score, and $Y_i$ is the actual outcome. If $R_i > T$, the patient's predicted outcome is 1, otherwise it is 0.

The performance of such a threshold-based classifier is often measured by sensitivity and specificity. In short, sensitivity represents the proportion of patients with $Y = 1$ that are correctly predicted as such, and specificity represents the proportion of patients with $Y = 0$ that are correctly predicted as such. A different classifier results in different values of sensitivity and specificity. Given a risk prediction model, the sensitivity and specificity values for all possible threshold-based classifiers can be summarized by the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). The AUC is a single value summary of the ROC curve.

The ROC curve is a two-dimensional graph in which the sensitivity (true positive rate) is plotted on the Y-axis. One minus the specificity (false positive rate) is plotted on the X-axis. We illustrate this in Figure 2.1. The coordinates of the origin indicate zero sensitivity and specificity of one, which is the case when $T = \infty$. The curve climbs upward on the Y-axis and rightward on the X-axis as the value of threshold deceases. Mathematically, the ROC curve plots the true positive rate as a function of $T$ versus the false positive rate as a function $T$. Let $f(\cdot)$ be the density function of $R$ when $Y = 1$, and $g(.)$ be the density function of $Y = 0$. For $R$, the true positive rate and the false positive rate can be expressed as $\int_T^\infty f(r)dr$ and $\int_T^\infty g(r)dr$, respectively. A patient will be classified as $Y = 1$ if its predicted risk of outcome is greater than $T$. Otherwise, it will be classified as $Y = 0$. Figure 2.1 illustrates the ROC curves of three models represented in green, red, and blue. The area under the ROC curves is the AUC. The larger the AUC, the better the predictive performance of the model. Thus, the model represented by the green curve reveals the best discriminatory ability among them all.

The c statistic is widely used discrimination measure of risk prediction models. It is equivalent to the AUC for dichotomous outcomes [15]. The definition of c

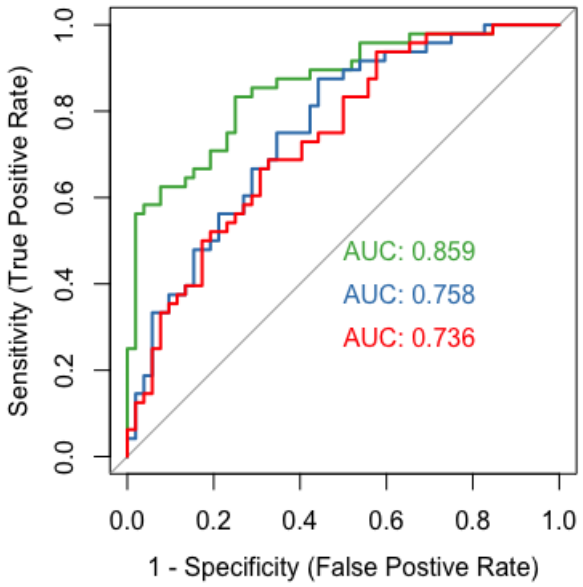



Figure 2.1: The ROC curves and values of AUC for three different sets of risk scores as an example. Plot using package 'pROC' in R [1]

statistic is based on the two-subject experiments, where two subjects are randomly selected to form a pair, and in the subject pair, one from the $Y = 1$ and the other from $Y = 0$ group. All pairs are assumed to be independent with each other. Each pair is made up of $\{(R_1, Y_1), (R_2, Y_2)\}$, where $(R_1, Y_1)$ are the risk score and the actual outcome of the first subject in a pair, and $(R_2, Y_2)$ are the corresponding attributes of the other subject.

With the two-subject experiments, the c statistic is defined as the probability of the event that, in a patient pair, one who experiences the outcome will have a higher predicted risk than the other who does not experience it. In other words, the c statistic of a predicted model is the concordance of $R$ and $Y$ given $Y_1 = 0$ and $Y_2 = 1$ [16] [17] [18] [11]. The $R$ is obtained from the model. Concordance, also known as the agreement, is closely related to correlation. Like correlation, it evaluates whether two variables tend to be large together or small together. In particular, $R$ and $Y$ are concordant if $\{R_1 < R_2, Y_1 < Y_2\}$ or $\{R_1 > R_2, Y_1 > Y_2\}$ holds [19] [20] [21]. Oppositely, $R$ and $Y$ are disconcordant if $\{R_1 < R_2, Y_1 > Y_2\}$ or $\{R_1 > R_2, Y_1 < Y_2\}$.

Altogether, the c statistic is expressed as

$$Pr\{R_2 > R_1 | Y_2 > Y_1\} = Pr\{R_2 > R_1 | Y_2 = 1, Y_1 = 0\}. \tag{2.2}$$

It is a popular measure of how well the risk predictions can discriminate the distributions of units with and without the outcome of interest.

### 2.2.2 The Lorenz curve and the Gini index

The Gini index (Gini coefficient) is a conventional measure of inequality primarily used in economics, which is originally defined by Corado Gini in 1921 [13]. The Gini index can be derived from the Lorenz curve, so we first discuss the Lorenz curve definition. Then, we will look more closely at the Gini index later in this subsection.

#### The Lorenz curve

The Lorenz curve was initially used in econometrics to describe the population's income inequality. The idea of the Lorenz curve can be adopted to explain the heterogeneity of the risk score distribution.

Recall that $R$ denotes a vector of risk scores from a risk prediction function. Let $F(r)$ and $f(r)$ be the cumulative distribution function and the probability density function of $R$, respectively. The Lorenz curve is first defined as a function of $R$:

$$L_1(r) = \frac{1}{\mu_r} \int_0^r t f(t) dt, \tag{2.3}$$

where $\mu_r$ represents the expected value of $R$. Let $p$ be the probability that $R$ is less than or equal to a given value $r$. Mathematically, $p = F(r)$. Based on this connection, $r$ in Equation 2.3 can be represented by the inverse of a cumulative distribution function of $p : r = F^{-1}(p)$ [22]. Thus, Equation 2.3 can be written as a function of probability $p$:

$$L_2(p) = \frac{1}{\mu_r} \int_0^p F^{-1}(t) dt, \tag{2.4}$$

where $L_2(p) = L_1(r)$.

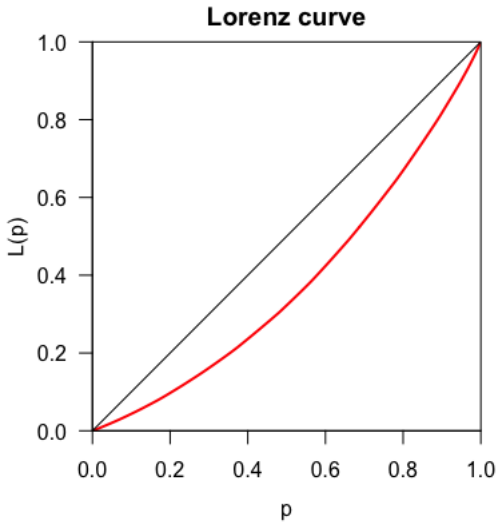



Figure 2.2: Lorenz curve example. Plot using package 'ineq' in R [2]

In Figure 2.2, the red curve represents $L_2(p)$ as a function of $p$, and the black line represents the perfect equality that the increasing rate of $p$ is identical to the increasing rate of $L_2(p)$. The perfect equality happens when everyone has the same risk, so the distribution of $R$ is a delta function. Any spread in $R$ and the Lorenz curve will be curved. The departure of the red curve from the black line reflects the difference between $L_2(p)$ and the perfect equality. Meanwhile, it is the departure of $L_1(p)$ from the perfect equality which implies the heterogeneity of the risk score distribution. The Gini index was developed to further quantify such inequality by summarizing the extent of heterogeneity.

### The Gini index

Using a geometric approach, the Gini index is defined as two times the area between the black line and the red curve shown in Figure 2.2 [23]. The area between the Lorenz curve and the straight line is also known as the area of concentration.

Thus, the Gini index can be expressed in the form of:

$$\begin{aligned} \text{Gini index} &= 2\int_0^1 (p - L(p))dp \\ &= 1 - 2\int_0^1 L(p)dp. \end{aligned} \tag{2.5}$$

An alternative definition for the Gini index is half of the relative mean difference. First of all, the absolute mean difference is the expected absolute difference between two realizations of *iid* random variables $R_1$ and $R_2$: $E[|R_1 - R_2|]$. The relative mean difference is defined as $E[|R_1 - R_2|]/\mu_r$, where $\mu_r$ is the expected value of $R$ [23]. Thus, the Gini index can be written as:

$$\text{Gini index} = \frac{E[|R_1 - R_2|]}{2\mu_r}. \tag{2.6}$$

The Lorenz curve and the Gini index visualize and quantify the heterogeneity of the distribution of risk score. However, one drawback of the above-mentioned definitions is that they are purely 'model-based' as they solely depend on the predicted risks not the observed outcomes. Therefore, we introduce a Gini-like index, which considers two variables and can be derived from the relative concentration curve. Note that this Gini-like index directly relates to one proposed metric in Chapter 3.

The relative concentration curve is a bivariate analogue of the Lorenz curve, and it summarizes the degree of association between distributions of a pair of variables [24] [25]. Here, the pair of variables of interest is $(R, Y)$. Thus, the relative concentration curve for $(R, Y)$ plots the cumulative proportion of $Y$ against $p = F(r)$.

In order to highlight the role of ranking variables [25], we rewrite the Lorenz curve $L(p)$ in Equation 2.3 as

$$L(p) = \frac{E[RI(R \leq F^{-1}(p))]}{\mu_r}.$$

We now consider another nonnegative random variable $Y$ with mean $\mu_y$. The relative concentration curve for $Y$ given $R$ is defined as:

$$L_{Y|R}(p) = \frac{E[YI(R \leq F^{-1}(p))]}{\mu_y},$$

where $R$ acts as a ranking mechanism, and it determines the summation order of computing the cumulative proportion of $Y$. Specifically, individuals with the risk score smaller than $F^{-1}(p)$ will be considered first.

The Gini-like index and the Gini index are defined similarly. The Gini-like index value equals two times the area of concentration in the relative concentration curve. In like manner, there is an alternative expression of the Gini-like index as the Gini index, which is expressed as half of the modified relative mean difference:

$$\text{Gini-like index} = \frac{E[Y_1 - Y_2 | R_1 > R_2]}{2\mu_y}. \tag{2.7}$$

Overall, the Gini-like index quantifies to what extent $Y$ and $R$ are associated with each other. There is a direct connection between the c statisitc and the Gini-like index, which is outlined in Section A.1.

## 2.3 Treatment benefit prediction models

In Section 2.1 and Section 2.2, we have discussed risk prediction models and how to quantify their predictive performance. In this section, we turn our attention to treatment benefit and the models used to predict the benefit.

In medical contexts, treatment benefit often refers to causal effect of the treatment. The term 'causal effect' is conceptualized based on counterfactual or potential outcomes. In the counterfactual framework, each unit has two potential binary responses: one is the $Y^{(1)}$ that would be observed under treatment, and the other is the $Y^{(0)}$ that would be observed under the placebo or generally under the control setting. The individual causal effect of the treatment can be defined as the algebraic difference between the counterfactual outcomes. In this thesis, we name the treatment's causal effect as the treatment benefit. Suppose the outcome is clinically unfavourable. Then the individual treatment benefit is expressed as

$$B = Y^{(0)} - Y^{(1)}, \tag{2.8}$$

and the treatment benefit takes value from $\{-1, 0, 1\}$ for a binary response. However, as the term "potential outcomes" emphasizes, we cannot observe two outcomes on the same unit simultaneously even if the two outcomes potentially exist [26]. Thus, the population-level average treatment benefit becomes the first quan-

tity to explore. It is defined as

$$E[B] = E[Y^{(0)} - Y^{(1)}]. \tag{2.9}$$

However, it is an overall summary of the treatment benefit within a population, and it provides no information about whether individuals are affected differently by the treatment. For instance, a positive $E[B]$ may be only attributable to a few patients with substantial benefits and a considerable proportion of patients with slight adverse effects. Treatment may only be beneficial for some patients with specific baseline characteristics and have no effect or even a harmful effect on others. Accordingly, learning the heterogeneity of treatment effect is a fundamental step in optimizing the effectiveness of the treatment. As it is impossible to measure $B$ individually, we intend to gain knowledge of heterogeneity from the distribution of the $E[B]$ measured from different subgroups. Each subgroup contains patients with the same value of $X$. Thus, the object of interest is a conditional expected treatment benefit in the form of

$$\begin{aligned} P &= E[B|X] \\ &= E[Y^{(0)}|X] - E[Y^{(1)}|X], \end{aligned} \tag{2.10}$$

where $P$ is a function of $X$.

In practice, the observed values from a randomized controlled trial are denoted as $(Y, X, A)$, where $A$ is a binary treatment assignment indicator. In this thesis, we consider an active treatment and control for $A = 1$ and $A = 0$, respectively. With respect to the observable data, an associational difference can be defined as

$$E[Y|X, A = 0] - E[Y|X, A = 1], \tag{2.11}$$

where $E[Y|X, A]$ reflects the association rather than causation between $A$ and $Y$. Equation 2.11 is identical to Equation 2.10 under the conditional exchangeability assumption which assumes that $(Y^{(0)}, Y^{(1)}) \perp\!\!\!\perp A|X$. Under the assumption, the causal quantity, $P$, can be identified by the associational difference, and the equivalence is shown as:

$$\begin{aligned} E[Y^{(0)}|X] - E[Y^{(1)}|X] &= E[Y^{(0)}|X, A = 0] - E[Y^{(1)}|X, A = 1] \\ &= E[Y|X, A = 0] - E[Y|X, A = 1]. \end{aligned}$$

The conditional exchangeability assumption holds conceptually as we consider the data from a randomized controlled trial (RCT). Recall that we aim to study a causal quantity. The randomized experimental design helps remove confounding effects

by having similar treatment group and control group distributions. However, the imbalances caused by chance may still remain. Thus, we consider relatively large sample sizes to avoid such nonstructural imbalance in practice [12].

For the binary response $Y$, a logistic regression model is commonly used to model $E[Y|X, A]$. Unlike Section 2.1, this logistic regression model contains a treatment indicator term and potentially several treatment interaction terms. Let $\theta$ be a column vector of model parameters, which is partitioned into three parts: $\theta^\tau = (\beta^\tau, \beta_\alpha, \beta_\Delta^\tau)$, where a column vector $\beta^\tau$ contains intercept and main covariates effects; a scalar $\beta_\alpha$ is the main treatment effect; and a column vector $\beta_\Delta^\tau$ contains treatment interaction effects. For the $i$-th patient, the model can be expressed as:

$$\text{logit}(E[Y_i = 1|X_i, A_i]) = X_i'\beta + \beta_\alpha A_i + A_i X_i \beta_\Delta, \;\; i = 1, 2, 3, ... \tag{2.12}$$

Let $X_i$ be a row vector, and $X_i'$ extends $X_i$ by adding 1 as the first element. Let $A_i$ be the treatment assignment indicator. Now, Equation 2.11 can be rewritten as

$$\begin{aligned} &E[Y_i|X_i, A_i = 0] - E[Y_i|X_i, A_i = 1] \\ &= \text{expit}\{X_i'\beta\} - \text{expit}\left\{X_i'\beta + \beta_\alpha A_i + A_i X_i \beta_\Delta\right\}, \end{aligned} \tag{2.13}$$

where the expit function is

$$\text{expit}(x) = \frac{1}{1 + \exp(-x)},$$

which is a monotonically increasing function of $x$.

Intuitively, if patients have various treatment effects on the medical outcome according to their corresponding covariates, the treatment effect is heterogeneous. Oppositely, if every patient experiences the same amount of treatment effect regardless of their covariates, the treatment effect is expected to be homogeneous. Whereas, for regression models with non-linear link functions such as logistic regression, no treatment interaction on the log-odds scale does not implies no interaction on the risk scale. Details are given as follows.

Let $\underline{\mathbf{0}}$ represent the zero vector. Equation 2.13 shows that if $\beta_\Delta \neq \underline{\mathbf{0}}$, then heterogeneous treatment effect is revealed on both log-odds scale and risk scale and vice versa. If both $\beta_\alpha = 0$ and $\beta_\Delta = \underline{\mathbf{0}}$, there is no treatment interaction on the risk scale as $E[Y|X, A' = \mathbf{0}] - E[Y|X, A' = I] = \underline{\mathbf{0}}$. If $\beta_\alpha \neq 0$ and $\beta_\Delta = 0$, there is no treatment interaction on the log-odds scale. However, the absence of inter-

action on the log-odds scale does not imply the absence on the risk scale. More specifically, for a linear regression model, zero treatment interaction indicates that one unit increase in $X$ leads to a constant change in associational difference on the risk scale. However, for a regression model with a non-linear link function, such as logit, probit, or tobit, one unit increase in $X$ on the log-odds scale (or whatever scale that a link function refers to) does not lead to a constant change on the risk scale. Thus, the interpretation of interaction is not straightforward for "non-linear" models.

## 2.4 The concordance statistic for benefit and the concentration of benefit

The concordance statistic for benefit $(cfb)$ and the concentration of benefit ($C_b$) are two metrics to measures the discriminatory performance of a model for treatment benefit, which conceptually extends the idea of the c statistic and the Gini index, respectively, from the risk to the treatment benefit domain.

The $cfb$ was first proposed by van Klaveren et al. [11]. The motivation for devising $cfb$ is that the conventional performance metrics, such as the c statistic, can only predict the risk of an outcome rather than treatment benefit. Thus, they aim to develop a metric capable of evaluating a treatment benefit model's performance. Let $P$ represent the expected treatment benefit given prognostic covariates $X$. In [11], $cfb$ is defined based on a given clinical trial and a multivariable model used to predict $P$, and the predicted $P$ is denoted as $\hat{P}$. In other words, this proposed metric depends on both sample and prediction models. We will expand the definition of $cfb$ and name this metric differently in Chapter 3. Additionally, the observed treatment benefit $\hat{B}$ is defined to estimate the unobserved $B$. We should note that $B$ is defined for each individual, but $\hat{B}$ is only defined for matched patient pairs, where each pair contains one treated patient and one control.

The calculation process for $cfb$ can be summarized into three steps:

1. Compute $\hat{P}$ for each individual based on a pre-determined model. Here, we consider logistic regression models for binary response variables.

2. Compute $\hat{B}$ and average $\hat{P}$ for each matched pair. Patients could be matched by different matching factors. In [11], the covariate matching and the predicted treatment benefit matching are considered. The predicted treatment

benefit matching is preferred as it is able to offer more similar patients in a matched pair. Hence, we mainly focus on the predicted treatment benefit matching procedure in this thesis.

3. Compute the c statistic for treatment benefit. Adopting the definition of c statistic detailed in Section 2.2.1, the $cfb$ can be interpreted as the concordance of $\hat{B}$ and $\hat{P}$ when considering possible pairs of matched pairs. The nuance of pairs of pairs is glossed over for now, and it will be explained in Chapter 3.

However, the $cfb$ at population-level is not clearly defined in [11] as the paper mainly focuses on introducing the calculation process for $cfb$. An assumption for $cfb$ was briefly mentioned in another paper [10]. There, "there is no correlation in the distribution of outcomes under the two treatments, conditional on the variables in the prediction model" is assumed. This assumption may be too strong to be true in practice [27].

The $C_b$ metric is proposed by Sadatsafavi et al. in [12]. Their original intention is to study the treatment effect heterogeneity on the decision scale. Thus, $C_b$ was devised to quantify how well the prognostic covariates can explain the heterogeneity. The heterogeneity of the covariate distributions sheds light on the differences in individual treatment benefits. To quantify how informative the knowledge of covariates is, $C_b$ is defined based on the comparison between two decision rules when the objective is to provide treatment to one of two randomly selected individuals. One is randomly providing the treatment to one of the two, and the other one is providing treatment to the subject with a higher predicted treatment benefit. The metric, $C_b$, quantifies the relative improvement in the efficiency of the latter decision rule compared with the former. Additionally, it is directly related to the Gini-like index discussed in Subsection 2.2.2. As such, $C_b$ extends the Gini-like index from the risk domain to the treatment benefit domain.

In [12], $C_b$ is clearly defined at the population-level, with parametric and semiparametric estimators given. To further study the properties of $cfb$ and $C_b$, Chapter 3 first clarifies the definition of the two metrics in the counterfactual framework and at the population-level based on an imposed predictor of treatment benefit. Then, the metrics' estimation procedures are discussed. In particular, the semiparametric version of $C_b$ will be mainly discussed.

# Chapter 3

# Metrics of treatment benefit

The motivations for developing $cfb$ and $C_b$ were briefly introduced in Section 2.4. This chapter presents more details about the proposed metrics. We expand the two proposed metrics to: $(cfb, C_b)$, $(cfb_h, C_{b,h})$, and $(\widehat{cfb}_h, \widehat{C}_{b,h})$. The notation $h$ represents a previously existing treatment benefit predictor, which we will comment upon later in Section 3.1.3. Both $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ are defined in the counterfactual outcome framework: $cfb$ and $C_b$ are population quantities, and $cfb_h$ and $C_{b,h}$ depend on $h$ and population. The last pair of metrics, $\widehat{cfb}_h$ and $\widehat{C}_{b,h}$, are estimators of $cfb_h$ and $C_{b,h}$, respectively.

In the discussion of this chapter, we use the same notation as Chapter 2. Recall that $Y^{(0)}$ and $Y^{(1)}$ are dichotomous counterfactual outcomes that would be observed under control and treatment, respectively. We assume Y is an untoward outcome, such that the treatment is beneficial for the $i$-th subject if $Y_i^{(0)} > Y_i^{(1)}$. In the counterfactual outcome framework, the population is characterized by its joint distribution of $(Y^{(0)}, Y^{(1)}, X)$, where $X$ is a set of prognostic covariates. With the population, the observed outcome for the $i$-th patient can be expressed as:

$$Y_i = Y_i^{(0)} I(A_i = 0) + Y_i^{(1)} I(A_i = 1), \;\; i = 1, 2, 3, ...$$

where $A_i = 0$ can be interpreted as assigning the placebo (or control) therapy to the $i$-th patients, and $A_i = 1$ represents assigning the treatment therapy to the $i$-th patients. We assume that the treatment is randomly assigned to patients.

## 3.1 Metrics in counterfactual outcome framework

In the counterfactual framework, the actual treatment benefit, $B = Y^{(0)} - Y^{(1)}$, and the expected conditional treatment benefit given the covariates, $P = E[B|X]$, are known. We also consider a two-subject design that contains all possible pairs of subjects from the population. A subject pair can be expressed as *iid* copies of $(Y_c^{(0)}, Y_c^{(1)}, X_c)$, where $c = 1, 2$. Since each subject in the population has $P$ and $B$, a subject pair can be expressed as $\{(P_1, B_1), (P_2, B_2)\}$. With dichotomous outcomes, $B$ is a categorical variable with three levels $\{-1, 0, 1\}$, which represent harm (-1), no effect (0), and benefit (1), respectively. $P$ is a function of $X$, where $X$ is assumed to have at least one continuous component such that $P$ is a continuous variable with $Pr(P_1 = P_2) = 0$.

### 3.1.1 Concordance statistics for benefit $(cfb)$

Based on the definition in Section 2.2.1, the concordance statistic for risk evaluates a risk prediction model's ability to discriminate a dichotomous outcome. Whereas, as we assume that $P$ and $B$ are available for individuals, no model is needed to predict the benefit. We aim to study the concordance between $P$ and $B$. In the two-subject experiment, $cfb$ is a conditional probability of $P_1 > P_2$ given $B_1 > B_2$. Thus, with the counterfactual outcome within the population, the concordance statistic for the benefit at the population-level is defined as:

$$cfb = Pr\{P_1 > P_2 | B_1 > B_2\}. \tag{3.1}$$

Alternatively, another definition of $cfb$ is based on the concept of the concordance with a constraint. That is,

$$cfb = Pr\{(P_1 - P_2)(B_1 - B_2) > 0 | B_1 \neq B_2)\}, \tag{3.2}$$

which shows that the probability of having a concordant pair given $B_1 \neq B_2$ gives the value of $cfb$. All two-subject pairs can be partitioned into three groups: the pairs with $B_1 > B_2$, the pairs with $B_1 < B_2$, and the pairs with $B_1 = B_2$. It is obvious that the constraint $B_1 \neq B_2$, shrinks the sample space, and we find that Equation 3.1 and Equation 3.2 are equivalent. (See Section A.2 for details.)

Equation 3.1 implies that many of the mathematical properties of the c statistic are also exhibited by $cfb$. First of all, its value is bounded by 0.5 and 1. Secondly, if $B$ and $P$ are independent, $cfb = 0.5$. In this case, the values of $P$ say nothing

about the values of $B$, and the knowledge of $X$ cannot provide any information about who will benefit the most from treatment. Mathematically, by the definition of independence and continuity of $P$, we have

$$\begin{aligned} cfb &= Pr\{P_1 > P_2 | B_1 > B_2\} \\ &= Pr\{P_1 > P_2\} \\ &= 0.5. \end{aligned}$$

Thirdly, if $B_1 > B_2$ always implies $P_1 > P_2$, then $cfb = 1$ showing the strongest possible monotonic association, or we can say, the knowledge of covariates helps evaluate treatment effect heterogeneity of the population data. Finally, if the association departs from independence but has not reached the perfect monotonic association, the value of $cfb$ falls between 0.5 and 1. The value of $cfb$ is close to 1, if a big proportion of concordant pairs in which the unit receiving greater $B$ also have greater $P$.

However, the value of $B$ is unobtainable in reality. Thus, definitions of $cfb$ are conditional on some unmeasurable quantities. Even if these definitions might be amorphous outside of the counterfactual outcome framework, they explicate the monotonic association of interest. In general, $cfb$ summarizes the ability of $X$ to describe the heterogeneity of treatment effect of the population data by the monotonic association. The larger the value of $cfb$, the stronger the ability of $X$.

### 3.1.2 Concentration of benefit $(C_b)$

Recall the definition of $C_b$ is based on two treatment decision rules, and the value of $C_b$ is calculated from a two-subject experiment. In this thesis, the $C_b$ is related to the semiparametric $C_b$ in [12]. Then, the population-level concentration of benefit is defined based on both $P$ and $B$ in the form of:

$$C_b = 1 - \frac{E[B]}{E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)]}, \tag{3.3}$$

where $E[B]$ is the expected actual treatment benefit under the randomized treatment assignment rule, and the denominator of the Equation 3.3 refers to the targeted treatment assignment rule applying the knowledge of $X$ embedded in $P$. The targeted treatment assignment rule always provides the treatment to the subject who has the higher $P$ in the two-subject experiment. The difference between the two decision rules is of interest. The difference is measured by the ratio of

the two expectations. If the targeted treatment assignment is way better than the randomized treatment assignment, the value of $C_b$ will be close to one.

There is another version of the Concentration of Benefit index discussed in [12], which is defined only based on $P$. These two versions are mathematically identical, as long as $P = E[B|X]$ holds. The key of the equivalence is shown as:

$$\frac{E[B]}{E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)]} = \frac{E[P]}{E[\max(P_1, P_2)]}. \tag{3.4}$$

See Section A.3 for details.

To simplify the notation, let

$$B^* = B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2).$$

It is convenient to explore the range of $C_b$ by taking advantage of the equivalence. Applying the Equation 3.4, we are able to get $E[B^*] - E[B] \geq 0$. It is the same as showing $E\left[\max(P_1, P_2)\right] - E[P] \geq 0$.

$$\begin{aligned} E\left[\max(P_1, P_2)\right] - E[P] &= E\left[\frac{P_1 + P_2 + |P_2 - P_1|}{2}\right] - E[P] \\ &= \frac{1}{2} \cdot (2E[P] + E\left[|P_2 - P_1|\right]) - E[P] \\ &= \frac{1}{2} E\left[|P_2 - P_1|\right], \end{aligned}$$

where $E\left[|P_2 - P_1|\right]$ is the expected value of a random variable taking non-negative values. Therefore, $E\left[|P_2 - P_1|\right] > 0$, which implies that $E[B^*] \geq E[B]$. Now, we discuss the possible values of $C_b$, which can be divided into four cases:

$$\begin{cases} 2 < C_b, & E[B] < 0 < E[B^*] < -E[B] \\ 1 < C_b \leq 2, & E[B] < 0 < -E[B] \leq E[B^*] \\ 0 \leq C_b \leq 1, & 0 \leq E[B] \leq E[B^*] \\ C_b \leq 0, & E[B] \leq E[B^*] < 0. \end{cases}$$

One main concern is that the value of $C_b$ is undefined if $E[B^*] = 0$. To get rid of the hindrance of the metric interpretation, the population-average treatment benefit is assumed to be strictly greater than zero ($E[B] > 0$). After making this assumption, the range of $C_b$ is $(0, 1]$.

Moreover, $C_b$ has the closed-form relationship with Gini index for benefit, which is defined as

$$Gini_b = \frac{E[B_1 - B_2 | P_1 > P_2]}{2E[B]}, \tag{3.5}$$

and the relationship between Equation 3.6 and Equation 3.3 can be expressed as

$$C_b = \frac{Gini_b}{Gini_b + 1}. \tag{3.6}$$

See Section A.5 for detail.

### 3.1.3 Concordance statistic for benefit and concentration of benefit with benefit predictors $(cfb_h, C_{b,h})$

In Subsection 3.1.1 and Subsection 3.1.2, we introduced population quantities $cfb$ and $C_b$. One must be mindful that the actual but unobservable relationship between covariates and treatment benefit is summarized by $P = E[B|X]$. To understand or 'learn' $P$, we consider a benefit predictor $h$, which is a function of $X$. The function $h(\cdot)$ can be developed in many ways. In this thesis, $h(\cdot)$ depends on a fully specified multivariable model with presumed model coefficients, and the model is externally defined. Note that when the function $h(\cdot)$ is fully developed, its external validation is of concern. Let $P_h$ be the predicted treatment benefit obtained from the function $h(\cdot)$. Mathematically, we have

$$P_h = h(X),$$

although $P_h$ might vary with only a subset of $X$.

We introduce an intermediate pair of metrics: $cfb_h$ and $C_{b,h}$, which depend on predicted treatment benefit $P_h$ and the population. Let $Y$ be an observed outcome in reality, and $A$ is the treatment assignment indicator. We assume that all observed data are from RCT to simplify the question.

This thesis considers binary responses. The logistic regression models are commonly used to model binary response variables. For instance, suppose we have a population of size $N$, and the imposed fully specified multivariable model is in the form of:

$$\text{logit}\left\{E[Y|X_1, X_2, A]\right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 A + \beta_4 A X_1,$$

where $\{X_1, X_2\} \subset X$. The predicted treatment benefit is expressed as:

$$P_h = \text{expit}\{\beta_0 + \beta_1 X_1 + \beta_2 X_2\} - \text{expit}\left\{(\beta_0 + \beta_3) + (\beta_1 + \beta_4) X_1 + \beta_2 X_2\right\}. \tag{3.7}$$

Similarly, $cfb_h$ and $C_{b,h}$ are defined based on two *iid* copies of $\{P_{hi}, B_i\}$, $i = 1, 2$. With the population and $P_h$, $cfb_h$ is defined as:

$$cfb_h = Pr\{P_{h1} > P_{h2} | B_1 > B_2\}, \tag{3.8}$$

and $C_{b,h}$ is defined as:

$$C_{b,h} = 1 - \frac{E[B]}{E[B_1 I(P_{h1} > P_{h2}) + B_2 I(P_{h1} < P_{h2})]}. \tag{3.9}$$

The benefit predictor may not always describe the underlying truth of the population. Only when $P_h = P$, we have $cfb_h = cfb$ and $C_{b,h} = C_b$. To put it another way, $cfb_h$ and $C_{b,h}$ are metrics for a presumed benefit predictor in the counterfactual outcome framework, which reflect the characteristics of both $h(\cdot)$ and the target population. Metrics, $C_b$ and $C_{b,h}$, quantify to what extent the knowledge of covariates helps describe the HTE. Such knowledge of covariates is embedded when applying the benefit predictor to get $P_h$. These two metrics imply the difference between the two treatment decision rules. One is applying treatment to whom has the highest $P$, and the other is randomly assigning the treatment.

According to the definition, $cfb$ is the monotonic association between $P$ and $B$. Thus, a large $cfb$ shows a stronger positive correlation between $P$ and $B$; in contrast, a smaller $cfb$ implies a weaker positive correlation. As patients with the same values of covariates have the same $P$, the variation of $X$ reflects the variation of $B$. In other words, both $cfb$ and $cfb_h$ measure how much the variation of $X$ reflects the HTE in the counterfactual outcome framework.

### 3.1.4 Stylized examples

To compare $cfb$ and $C_b$, first we consider a simple linear set-up for which these metrics have closed forms that can be illustrative of their basic properties. Here, we consider two versions with the one in which counterfactual outcomes are independent conditional on predictors and the other in which counterfactual outcomes are dependent. In the first version, $P$ and $B$ are two linear functions of a covariate.

The purpose is to explore how $cfb$ and $C_b$ would be influenced by the values of the main treatment effect and the treatment interaction. Later, dependent counterfactual outcomes are introduced to the set-up. Then we can explore the correlation between potential outcomes and study whether the values of the metrics are influenced by the correlation.

**The version with independent counterfactual outcomes**

In this section, $X$ is a covariate generated from standard normal distribution, and the expected treatment benefit $P$ and actual treatment benefit $B$ are structured directly as:

$$\begin{aligned} P &= \beta_0 + \beta_1 X, \\ B &= \beta_0 + \beta_1 X + \varepsilon \\ &= P + \varepsilon, \end{aligned}$$

where $\varepsilon$ (error) is a random variable generated from $N(0, \sigma^2)$, $\beta_0$ represents the main treatment effect, and $\beta_1$ is the treatment interaction effect. According to the set-up, $\varepsilon$ quantifies the deviation of $P$ from $B$. Since both $P$ and $B$ are continuous quantities, there are no ties that need to be adjusted.

Based on two-subject experiment and Equation 3.1, we have

$$\begin{aligned} cfb &= Pr\{P_1 > P_2 | B_1 > B_2\} \\ &= \frac{Pr\{P_1 > P_2, B_1 > B_2\}}{Pr\{B_1 > B_2\}} \\ &= 2Pr\{P_1 > P_2, B_1 > B_2\} \\ &= 2Pr\{P_2 - P_1 < 0, B_2 - B_1 < 0\}, \end{aligned} \tag{3.10}$$

where $((P_2 - P_1), (B_2 - B_1))^T$ has a bivariate normal distribution with mean $\mu$ and variance covariance matrix $\Sigma$. In this case,

$$\begin{aligned} \mu &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \Sigma &= \begin{pmatrix} 2\beta_1^2 & 2\beta_1^2 \\ 2\beta_1^2 & 2(\beta_1^2 + \sigma^2) \end{pmatrix}. \end{aligned}$$

Let $F_{P_2-P_1, B_2-B_1}\left((0,0)^T; \mu, \Sigma\right)$ be the cumulative distribution of the joint nor-

mal. Equation 3.10 states that

$$cfb = 2F_{P_2-P_1,B_2-B_1}\left((0,0)^T;\mu,\Sigma\right). \tag{3.11}$$

In words, the value of $cfb$ is equal to two times the probability that both components of $(P_2 - P_1, B_2 - B_1)$ are less than zero. (See Section A.6.1 for details.)

The value of $C_b$ can be calculated according to Equation 3.4, where $E[P] = \beta_0$, and

$$\begin{aligned}
&E[\max\{P_1, P_2\}] \\
&= E[\beta_0 + \beta_1 \max\{X_1, X_2\} I(\beta_1 \geq 0) + \beta_1 \min\{X_1, X_2\} I(\beta_1 < 0)] \\
&= \beta_0 + \beta_1 \left( \frac{1}{\sqrt{\pi}} I(\beta_1 \geq 0) - \frac{1}{\sqrt{\pi}} I(\beta_1 < 0) \right).
\end{aligned}$$

Then, $C_b$ can be expressed as:

$$C_b = \begin{cases} 1 - \frac{\beta_0}{\beta_0 + \frac{\beta_1}{\sqrt{\pi}}} = \frac{1}{\sqrt{\pi}\frac{\beta_0}{\beta_1} + 1} & \beta_1 \geq 0 \\ 1 - \frac{\beta_0}{\beta_0 - \frac{\beta_1}{\sqrt{\pi}}} = \frac{1}{-\sqrt{\pi}\frac{\beta_0}{\beta_1} + 1} & \beta_1 < 0. \end{cases}$$

(See Section A.7 for details.) If the value of $\beta_1$ is a constant, then $C_b$ is a reciprocal function of $\beta_0$. According to the positive population-average benefit assumption, we have $\beta_0 > 0$.

Figure 3.1 shows $cfb$ and $C_b$ as functions of $\beta_0$, or $\beta_1$, or $\sigma^2$. Each plot fixes two variables and let the third one vary. For example, when $\beta_0$ changes from 0 to 10, we have $\beta_1 = \sigma^2 = 1$. When $\beta_1$ varies from -10 to 10, $\beta_0 = \sigma^2 = 1$.

The top left plot shows the value of $C_b$ decreases as the absolute value of the main treatment effect increases, but the value of $cfb$ stays the same, as it is not a function of $\beta_0$. The plot on the top right shows the value of $cfb$ and $C_b$ increase when the absolute treatment interaction value increases, where $C_b$ increases faster than $cfb$. The plot on the bottom shows the relationships between metrics and the error, and the error is the difference between $B$ and $P$ by subtracting them. The plot shows that the value of $cfb$ decreases once the variation of the error increases. This time, $C_b$ is not influenced by the variance of error as it is a ratio of two expectations.

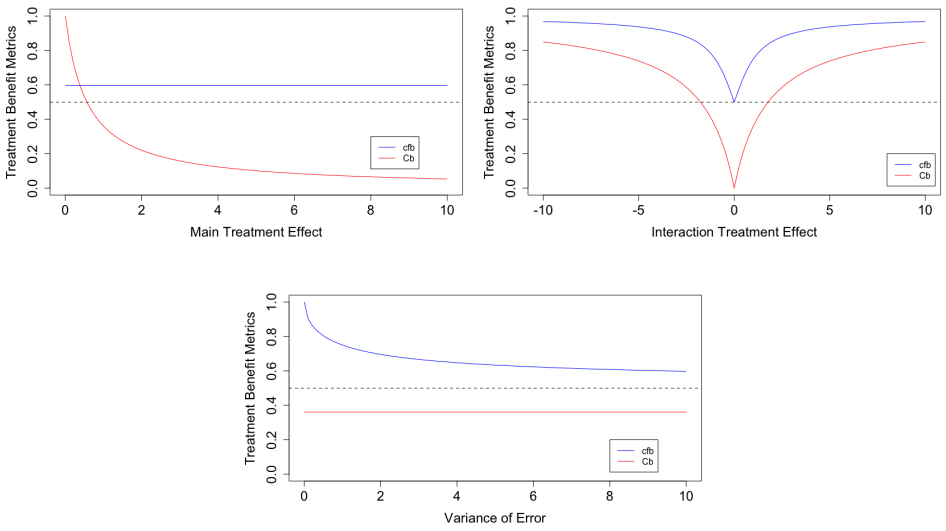



Figure 3.1: Relationships between main treatment effect, treatment interaction effect, variance of error and $(cfb, C_b)$.

### The version with dependent counterfactual outcomes

We adopt the notation from the previous subsection and modify the set-up by introducing counterfactual outcomes $Y^{(0)}$ and $Y^{(1)}$. We assume that

$$Y^{(0)} = (\alpha_0 + \beta_0) + (\alpha_1 + \beta_1)X + \varepsilon_0,$$
$$Y^{(1)} = \alpha_0 + \alpha_1 X + \varepsilon_1,$$

where $\varepsilon_0$ and $\varepsilon_1$ are potentially correlated random variables following the normal distribution with mean 0 and variance $\sigma^2$. Note that $\varepsilon_0$ and $\varepsilon_1$ are independent of $X$, respectively, and they could have different distribution parameters or from two different distributions in a more general case. Let $\beta_0$ be the main treatment effect and $\beta_1$ be the interaction treatment effects. As $Y^{(0)}$ and $Y^{(1)}$ are linear combinations of normal distributions, we have:

$$Y^{(0)} \sim N(\alpha_0 + \beta_0, (\alpha_1 + \beta_1)^2 + \sigma^2),$$
$$Y^{(1)} \sim N(\alpha_0, \alpha_1^2 + \sigma^2).$$

The correlation between $Y^{(0)}$ and $Y^{(1)}$ is not identifiable in practice as it cannot be measured. The current set-up gives a intuitive insight into the correlation:

$$\begin{aligned} cov(Y^{(0)}, Y^{(1)}) &= cov\left((\alpha_0 + \beta_0) + (\alpha_1 + \beta_1)X + \varepsilon_0, \alpha_0 + \alpha_1 X + \varepsilon_1\right) \\ &= cov((\alpha_1 + \beta_1)X, \alpha_1 X) + cov(\varepsilon_0, \varepsilon_1) \\ &= (\alpha_1 + \beta_1)\alpha_1 Var(X) + cov(\varepsilon_0, \varepsilon_1). \end{aligned}$$

The covariance of $Y^{(0)}$ and $Y^{(1)}$ is made up of two components with the one arising from the variance of $X$ and the other being the covariance of two error terms. Let $\rho$ be the correlation coefficient of $\varepsilon_0$ and $\varepsilon_1$. Note that since $X$ is observable, the first component of the covariance is identifiable, but $cov(\varepsilon_0, \varepsilon_1)$ is always left unrevealed, unless the fundamental problem of causal inference is solved. In this case, we have

$$\begin{aligned} cov(\varepsilon_0, \varepsilon_1) &= \rho\sigma^2, \\ cov(Y^{(0)}, Y^{(1)}) &= (\alpha_1 + \beta_1)\alpha_1 + \rho\sigma^2. \end{aligned}$$

The bivariate normal distribution of $Y^{(0)}$ and $Y^{(1)}$ can be expressed as:

$$\begin{pmatrix} Y^{(0)} \\ Y^{(1)} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_0 + \beta_0 \\ \alpha_0 \end{pmatrix}, \begin{pmatrix} (\alpha_1 + \beta_1)^2 + \sigma^2 & (\alpha_1 + \beta_1)\alpha_1 + \rho\sigma^2 \\ (\alpha_1 + \beta_1)\alpha_1 + \rho\sigma^2 & \alpha_1^2 + \sigma^2 \end{pmatrix}\right).$$

Now, based on the definitions of $P$ and $B$, we have

$$\begin{aligned} B &= Y^{(0)} - Y^{(1)} \\ &= \beta_0 + \beta_1 X + \varepsilon_0 - \varepsilon_1, \\ P &= E[B|X] \\ &= \beta_0 + \beta_1 X, \end{aligned}$$

where $B \sim N(\beta_0, \beta_1^2 + 2\sigma^2(1-\rho))$ and $P \sim N(\beta_0, \beta_1^2)$. In like manner, $cfb$ is expressed as

$$cfb = 2F_{P_2 - P_1, B_2 - B_1}\left((0,0)^T; \mu, \Sigma\right), \tag{3.12}$$

where the mean and the covariance matrix are

$$\begin{aligned} \mu &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \Sigma &= \begin{pmatrix} 2\beta_1^2 & 2\beta_1^2 \\ 2\beta_1^2 & 2\beta_1^2 + 4(1-\rho)\sigma^2 \end{pmatrix}. \end{aligned}$$

$C_b$ is expressed as

$$C_b = \frac{\beta_1}{\sqrt{\pi}\beta_0 + \beta_1}. \tag{3.13}$$

(See Section A.6.2 and Section A.7 for details.)

Comparing the results from these two versions of the linear set-up, we find that the value of $cfb$ is influenced by the hidden correlation $\rho$, but the value of $C_b$ remains the same, regardless of the size of $\rho$.

## 3.2 Metric estimators

Out of the counterfactual framework, we are not able to observe both $Y^{(0)}$ and $Y^{(1)}$ at the same time, and only the sample data are available for analysis. We assume that the sample data are a simple random sample from the population joint distribution of $(Y, X, A)$, where $X$ is a set of covariates.

Without observing the counterfactual outcomes, the exact value of individual $B$ is not available. Then, $P$ cannot be obtained by taking the conditional expectation of $B$ on $X$. Therefore, in this section, we focus on $P_h$, which refers to pre-defined benefit models. Then, we introduce estimation methods for $cfb_h$ and $C_{b,h}$.

The first step of finding the estimators for $cfb_h$ and $C_{b,h}$ is to calculate $P_h$ according to Equation 3.7. Note that Equation 3.7 emphasizes that as long as $h(X)$ is specified for the population, $P_h$ is immediately determined for the data. In this section, we use an externally imposed $h(X)$. The only difference is that $P_h$ in Section 3.1.3 is calculated using the population, and $P_h$ Section 3.2 is calculated using samples from the population.

The second steps for calculating $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ are quite different as discussed in Section 3.2.1 and Section 3.2.2.

### 3.2.1 Estimator of concordance statistic for benefit with benefit predictors ($\widehat{cfb_h}$)

The next step for $\widehat{cfb_h}$ is to estimate individual $B$ by the observed treatment benefit $\hat{B}$, which is defined as the outcome difference of a matched pair. Two possible

matching methods for creating such pairs are discussed in [11]: one is matching on predicted benefit $P_h$, and another one matching directly on covariates $X$.

The basic idea of estimating $B$ is using a matching method to adjust the distribution of predicted benefit $P_h$ or covariates $X$ in treated and control arms. Furthermore, each matched pair contains a patient from the treatment arm and a patient from the control arm. In a matched pair, two patients are expected to have the same $P_h$. Then, $\hat{B}$ is calculated as the outcome difference of these two patients with the same $P_h$. On the other hand, if two patients are matched by $X$, then $\hat{B}$ is defined as the outcomes difference of two patients with the same $X$. As $P_h$ is a function of $X$, we mainly discuss using the $P_h$ as the matching factor. Since it is almost impossible to find a pair of patients with the exact same matching factor in practice, the matched pair is formed by two patients with the most similar matching factor. We will comment upon the matching algorithm in Chapter 4.

The $\widehat{cfb}_h$ is expressed as

$$\widehat{cfb}_h = Pr\{\tilde{P}_{h,1} > \tilde{P}_{h,2} | \hat{B}_1 > \hat{B}_2\},$$

where $\tilde{P}_h$ is the average $P_h$ of two paired subjects, and $\hat{B}$ is the observed treatment benefit. Figure 3.2 demonstrates the value of $cfb$ or the value of $cfb_h$ is defined based on a pair of subjects, and each subject in the population has either $(P, B)$ or $(P_h, B)$. However, $\widehat{cfb}_h$ is defined based on a pair of pairs, which refers to four subjects. Each subject has $(P_h, Y)$. Subjects are paired to obtain $\tilde{P}_h$ and $\hat{B}$. Finally, the subject pairs are paired again to calculate $\widehat{cfb}_h$.

Overall, the calculation of $\widehat{cfb}_h$ has two steps. The first step is computing $P_h$ for each individual and forming matched pairs. Then, $\hat{B}$ and $\tilde{P}_h$ are calculated for each matched pair. The second step is to determine the value of $\widehat{cfb}_h$ by quantifying the concordance statistic among the pair of pairs.

### 3.2.2 Estimator of concentration of benefit with benefit predictors ($\widehat{C}_{b,h}$)

With the predicted treatment benefit $P_h$, the next step for $\widehat{C}_{b,h}$ is to rank all patients in the sample by their corresponding $P_h$. A chain of subgroups is constructed according to rank. The first subgroup contains the patient with the largest $P_h$ among all patients. The size of the second subgroup increases by 1, and it contains the

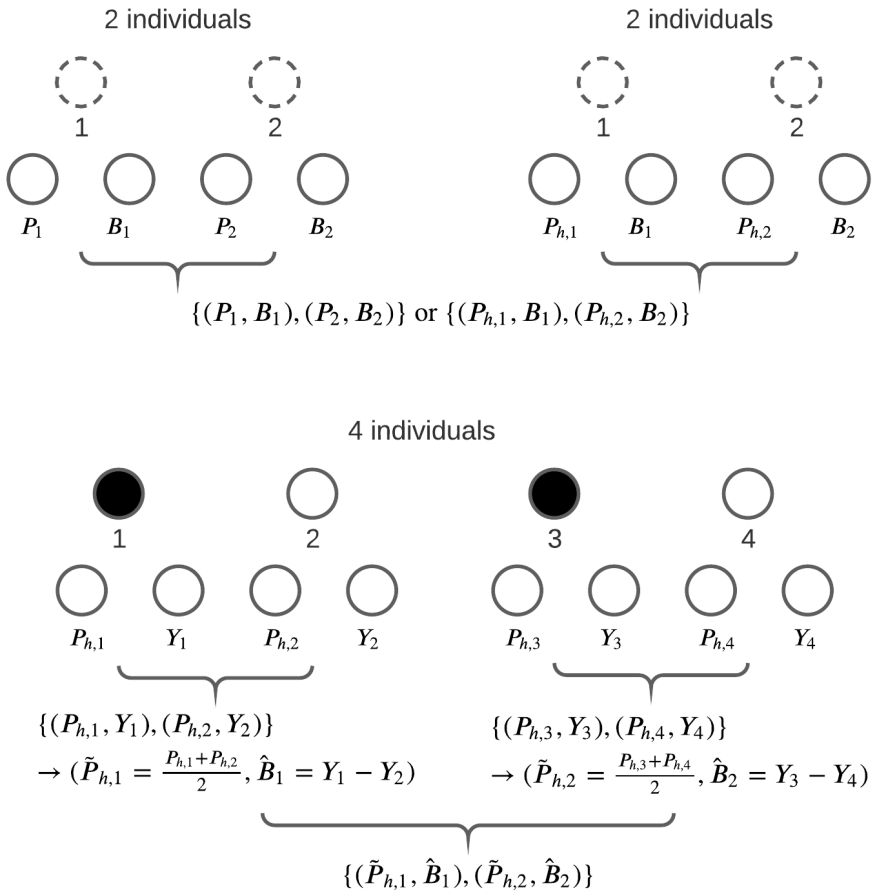



Figure 3.2: Two-subject experiment versus pair of pairs. Two subjects are used to calculate the value of $\widehat{cfb}$ or the value of $cfb_h$. Four subjects are required for calculating the value of $\widehat{cfb}_h$, where black circles with numbers 1 and 3 represent subjects from the control arm and white circles with numbers 2 and 4 represent subjects from the treatment arm.

previous patient and the patient with the second highest $P_h$ among all. By analogy, the $r$-th subgroup contains the first $r$ patients with the highest $P_h$, and the $(r+1)$-th subgroup contains the first $r+1$ patients with the highest $P_h$.

To calculate $\widehat{C}_{b,h}$, we need to estimate $E[B]$ and $E[B_1 I(P_{h,1} > P_{h,2}) + B_2 I(P_{h,1} < P_{h,2})]$, but before introducing the approach we first discuss the proposed approach used to estimate $E[P_h]$ and $E[\max(P_{h,1}, P_{h,2})]$ in [12].

First of all, when the population is finite with size $N$, there are finite number of two-subject pairs. Then, the equivalence that

$$E[\max(P_{h,1}, P_{h,2})] = \frac{1}{N^2}\left\{2\sum_{r=1}^{N}\hat{S}(r) - \frac{1}{N}E[P_h]\right\} \tag{3.14}$$

can be proved, where $\hat{S}(r)$ is cumulative predicted benefit. That it, the partial sum of the predicted benefit of the $r$-th subgroup. (See Section A.4 for details.)

Now, consider a finite sample data set of size n. The expectation, $E[\max(P_{h,1}, P_{h,2})]$, is estimated as

$$\hat{E}[\max(P_{h,1}, P_{h,2})] = \frac{1}{n^2}\left\{2\sum_{r=1}^{n}\hat{S}(r) - \frac{1}{n}\hat{E}[P_h]\right\}.$$

Note that $\hat{E}[P_h]$ as the estimator of $E[P_h]$ is a empirical mean of observations of $P_h$. We write this as:

$$\hat{E}[P_h] = \frac{1}{n}\sum_{r=1}^{n} p_{h,r},$$

where $p_{h,r}$ is the realization with the rank $r$. In the similar manner of Equation 3.14, $E[B_1 I(P_{h,1} > P_{h,2}) + B_2 I(P_{h,1} < P_{h,2})]$ is estimated as

$$\hat{E}[B_1 I(P_{h,1} > P_{h,2}) + B_2 I(P_{h,1} < P_{h,2})] = \frac{1}{n^2}\left\{2\sum_{r=1}^{n}\widehat{S}_B(r) - \frac{1}{n}\hat{E}[B]\right\}, \tag{3.15}$$

where $\widehat{S}_B(r)$ is the cumulative $B$ of the $r$-th subgroup. Note, however, that because only one potential outcome corresponding to the treatment assignment is available, the actual $B$ is unobservable. Thus, the $E[B]$ is estimated by the average observed outcome difference between two treatment arms (Neyman difference-in-means es-

timator). That is

$$\hat{E}[B] = \frac{\sum_{i=1}^{n} I(A_i = 0)Y_i}{\sum_{i=1}^{n} I(A_i = 0)} - \frac{\sum_{i=1}^{n} I(A_i = 1)Y_i}{\sum_{i=1}^{n} I(A_i = 1)}. \tag{3.16}$$

Similarly, the cumulative $B$ of the $r$-th subgroup can be estimated as

$$\hat{S_B}(r) = r \left( \frac{\sum_{i=1}^{r} I(A_i = 0)Y_i}{\sum_{i=1}^{r} I(A_i = 0)} - \frac{\sum_{i=1}^{r} I(A_i = 1)Y_i}{\sum_{i=1}^{r} I(A_i = 1)} \right). \tag{3.17}$$

The last step is to combine Equation 3.15, Equation 3.16 and Equation 3.17 to get the estimate of $C_{b,h}$ is

$$\widehat{C}_{b,h} = 1 - \frac{\hat{E}[B]}{\hat{E}[B_1 I(P_{h,1} > P_{h,2}) + B_2 I(P_{h,1} < P_{h,2})]}. \tag{3.18}$$

To sum up, it is not necessary to compute $\hat{E}[B]$ by estimating each individual's actual treatment benefit. Instead, Equation 3.16 shows that $E[B]$ is estimated by the difference in the sample's average outcome between treatment groups. The estimator is unbiased under the assumption that treatment assignment $A$ and counterfactual outcomes $Y^{(0)}$ and $Y^{(1)}$ are independent given the covariates $X$ [28]. Otherwise, bias arises. Similarly, Equation 3.17 can provide biased results if the conditional independence assumption is not satisfied.

# Chapter 4

# Simulation studies

In this chapter, we investigate and compare $(cfb, C_b)$, $(cfb_h, C_{b,h})$, and $(\widehat{cfb}_h, \widehat{C}_{b,h})$ in various settings, with notation adopted from Chapter 3. We begin with an overview of the algorithms used to compute the metrics. In subsequent sections, we construct simulations where outcome variable $Y$ is binary, and a set of prognostic covariates $X$ contains at least one continuous variable. The values of metrics from populations with different HTE will be explored in Section 4.2. The simulations in Section 4.3 demonstrate the differences between metrics when counterfactual outcomes $Y^{(0)}$ and $Y^{(1)}$ are not independent and when the covariates are pair-wise correlated.

Before introducing the simulation studies, it is helpful to recall the metrics' definitions in Chapter 3. Both $cfb$ and $C_b$ summarize the underlying treatment benefit properties of the population without considering any models, and they can be computed only if the covariates of interest and the corresponding binary counterfactual outcomes $Y^{(0)}$ and $Y^{(1)}$ are available. Then, $cfb_h$ and $C_{b,h}$ are another pair of metrics discussed under the counterfactual framework. These two are defined based on the treatment benefit predictor $h(\cdot)$ that uses the knowledge of prognostic covariates $X$. The types of $h(\cdot)$ an vary, but we only discuss the $h(\cdot)$ referring to externally imposed function of $X$. Lastly, $(\widehat{cfb}_h, \widehat{C}_{b,h})$ represent the estimators of $cfb_h$ and $C_{b,h}$ respectively. For simplicity, we assume that data are collected from a complete RCT, where subjects are randomly assigned with a probability of 0.5 to the treatment and control arm.

## 4.1 Algorithms for metrics computation

The algorithms for $(cfb, C_b)$, $(cfb_h, C_{b,h})$ and $(\widehat{cfb}_h, \widehat{C}_{b,h})$ are implemented in R.

We calculate $cfb$ and $C_b$ based on the expected treatment benefit $P$ and the actual treatment benefit $B$. As we know that $cfb$ is an analogue of the c statistic, it can be computed using the 'rcorr.cens' function in the Hmisc package [29]. The 'rcorr.cens' function takes two inputs with the one being a numeric predictor variable and the other being the response variable of interest. Thus, to compute the c statistic of treatment benefit, $P$ and $B$ are taken as two inputs of the function.

The definition of $C_b$ is one minus the ratio of $E[B]$ and $E[B^*]$, where $E[B]$ can be calculated with ease. One difficulty is computing the value of $E[B^*]$, where $B^* = B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)$. To compute $E[B^*]$ efficiently, the corresponding algorithm is not implemented based on two-subject experiments as defined. Instead, it is calculated according to the similar trick discussed in Equation 3.15, but with the actual $E[B]$ and $S_B(k)$. In the algorithm, $B$ is sorted into descending order by $P$. After we get $E[B]$ and $E[B^*]$, the values are plugged into Equation 3.3 to compute $C_b$.

To calculate $cfb_h$ and $C_{b,h}$, we need to obtain $P_h$ using imposed models. As we consider a binary response variable in this thesis, $P_h$ is the difference of two predictions $\hat{E}[Y|X, A = 0]$ and $\hat{E}[Y|X, A = 1]$ based on a predetermined logistic regression model. With the inputs $P_h$ and $B$, $cfb_h$ and $C_{b,h}$ can be computed using the above algorithms.

Since it is impossible to obtain $B$ and the data of an entire population in practice, it is important to consider approaches to compute $\widehat{cfb}_h$ and $\widehat{C}_{b,h}$. Now, we obtain $P_h$ by sample data, and the actual treatment benefit $B$ is estimated by the $\hat{B}$ which is the outcomes difference of a matched pair with one subject from the treatment arm and the other from the control arm.

The algorithm for $\widehat{cfb_h}$ can be divided into a calculation step and an iteration step. The estimation step is to obtain $P_h$ and $\hat{B}$, and the iteration step is to compute the concordance statistic of $P_h$ and $\hat{B}$. This is achieved by counting the number of concordant pairs of pairs and calculating the probability of the concordant pairs.

- *Calculation step:* With an imposed model and observed sample data, $P_h$ is calculated by the difference between $\hat{E}[Y = 1|X, A = 0]$ and $\hat{E}[Y = 1|X, A = 1]$ using the imposed model. Recall that $\hat{B}$ is defined based on a

pair of matching units who have the same $P_h$ but in different treatment arms [11]. It is important to clarify the definition of similarity, a point we return to later in this section. Therefore, each pair is supposed to have a corresponding $\hat{B}$ and a similar (ideally identical) $P_h$.

- *Iteration step:* Two pairs of units are randomly drawn from all matching pairs, where all pairs are independent of each other. We first initialize the numerator to zero and set the denominator as the total number of two pairs combinations. Then, we loop through all possible combinations. The $i$-th iteration step is checking whether $P_h$ and $\hat{B}$ of two pairs of units (pair 1 and pair 2) are concordant. That is, whether

$$(\hat{B}_1 - \hat{B}_2)(P_{h,1} - P_{h,2}) > 0.$$

  If so, add one to numerator. If $\hat{B}_1 - \hat{B}_2 = 0$ or $P_{h,1} - P_{h,2} = 0$, exclude this combination by subtracting one from the denominator. Overall, $cfb$ is calculated as the proportion of concordant pairs over all valid pair of pairs.

Similarly, to estimate $C_{b,h}$, all sample subjects are sorted in descending order by $P_h$. Then, we use the average observed outcome difference between two treatment groups (as Equation 3.16) to estimate $E[B]$, and $E[B^*]$ is estimated by applying Equation 3.15. Finally, the estimations are substituted in Equation 3.18 to obtain $\widehat{C}_{b,h}$.

### Matching of patients

It is hard to find two patients with the same $P_h$, noting that $P_h$ is continuous in practice. The matching procedure influences the value of $\hat{B}$ and $\widehat{cfb}_h$. The sensitivity to matching procedure choice should be explored in the future. In this thesis, the similarity (or say distance) between patients is assessed by a pair-wise difference of $P_h$. A greedy approach is performed to pair a patient from the treatment group to another from the control group. A match is chosen for each treated unit one at a time. In each matching step, we choose the control unit that is not yet matched but is the closest to the current treated unit. Patients who are not paired will be dropped from the sample.

Such a method can be implemented through MatchIt [30] which is an R package. Here, we use the predicted treatment benefit difference as distance and apply the nearest distance matching method to form matching pairs.

## 4.2 Simulation studies with different population HTE

In this section, simulations are performed to study how much heterogeneity of treatment effect (HTE) in population data has influence on the values of the three pairs of metrics: $(cfb, C_b)$, $(cfb_h, C_{b,h})$ and $(\widehat{cfb}_h, \widehat{C}_{b,h})$.

### 4.2.1 Simulation setup

Suppose the joint distribution of the target population is $(Y^{(0)}, Y^{(1)}, X)$, where $X$ is a collection of prognostic variables, and $Y^{(0)}$ and $Y^{(1)}$ are assumed to be independent. The true benefit is defined as the counterfactual outcomes difference $B = Y^{(1)} - Y^{(0)}$. The population's characteristic is summarized by the conditional expectation of the benefit given a set of prognostic variables, denoted as $g(X) = P = E[B|X]$.

We consider a set of observed prognostic variables $X = \{X_1, X_2, X_3\}$ generated as *iid* standard normal continuous variables. The logistic regressions can be described as

$$\text{logit}\left(E[Y^{(0)}|X_1, X_2, X_3]\right) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \tag{4.1}$$

$$\begin{aligned}\text{logit}\left(E[Y^{(1)}|X_1, X_2, X_3]\right) = (\beta_1 + \beta_{1,\Delta})X_1 + (\beta_2 + \beta_{2,\Delta})X_2 + \\ (\beta_3 + \beta_{3,\Delta})X_3 + \beta_a,\end{aligned} \tag{4.2}$$

where the $\beta_{i,\Delta}, i = 1, 2, 3$ represents the interaction of treatment, and $\beta_a$ represents the main treatment effect. Note that $\beta_{2,\Delta}, \beta_{3,\Delta}$ are assumed to be zero in this chapter. Thus, if $\beta_{1,\Delta} = 0$, all individuals who have received the treatment will have the same amount of treatment effect on the log odds scale. By contrast, the larger value of $\beta_{1,\Delta}$, the more remarkable the interpersonal differences in terms of the treatment benefit.

We simulate a population of size $10^5$, and let $\beta_1 = 1, \beta_2 = 0.5, \beta_3 = -0.4$. We assume $\beta_{1,\Delta} \in \{0, -0.05, -1, -4\}$ and $\beta_a \in \{-2, -6\}$. Thus, there are eight different combinations of $\beta_{1,\Delta}$ and $\beta_a$, which refer to eight populations. Populations appear in the Table 4.1.

Returning to Chapter 3, we know that each population has unique $C_b$ and $cfb$ values, which only depend on the actual relationship defined in Equation 4.1 and Equation 4.2. To put it another way, $C_b$ and $cfb$ reflect the extent to which covari-

| *Population* | $\beta_1$ | $\beta_{1,\Delta}$ | $\beta_2$ | $\beta_3$ | $\beta_a$ |
|---|---|---|---|---|---|
| A | 1 | 0 | 0.5 | -0.4 | -2 |
| B | 1 | -0.05 | 0.5 | -0.4 | -2 |
| C | 1 | -1 | 0.5 | -0.4 | -2 |
| D | 1 | -4 | 0.5 | -0.4 | -2 |
| E | 1 | 0 | 0.5 | -0.4 | -6 |
| F | 1 | -0.05 | 0.5 | -0.4 | -6 |
| G | 1 | -1 | 0.5 | -0.4 | -6 |
| H | 1 | -4 | 0.5 | -0.4 | -6 |

Table 4.1: Eight populations with sets of model coefficients

ates explain the HTE of the population.

We compute $C_{b,h}$ and $cfb_h$ based on the population, random treatment assignment $A$ and $h(\cdot)$. For each population, we consider five benefit predictors, which are denoted as $h = 1, 2, 3, 4, 5$ and related to five fully specified logistic regression models. Model structures for these five models are shown as follows:

1. $Y \sim X_1 + X_2 + X_3 + A$,
2. $Y \sim X_1 + X_2 + X_3 + A + X_1 : A + X_2 : A + X_3 : A$,
3. $Y \sim X_1 + A + X_1 : A$,
4. $Y \sim X_2 + A + X_2 : A$,
5. $Y \sim X_3 + A + X_3 : A$.

The unknown model parameters are estimated by the fitting the corresponding model to the population. The models' parameter estimates for eight population data are shown in the Table B.1 and Table B.2.

Now, we estimate $C_{b,h}$ and $cfb_h$ using sample data, and the estimators are denoted as $\widehat{C}_{b,h}$ and $\widehat{cfb_h}$, respectively. Consider the observed sample data $(Y, X, A)$. We simulated 200 sample data sets for every population, and each data set contains $n = 500$ observations.

Table 4.2: The values of $(C_b, cfb)$ and $(C_{b,h}, cfb_h)$ for eight populations with different HTE

| *Population* | $C_b$ | $C_{b,h}$ | | | | | $cfb$ | $cfb_h$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $C_{b,1}$ | $C_{b,2}$ | $C_{b,3}$ | $C_{b,4}$ | $C_{b,5}$ | | $cfb_1$ | $cfb_2$ | $cfb_3$ | $cfb_4$ | $cfb_5$ |
| A | 0.167 | 0.167 | 0.167 | 0.135 | 0.062 | 0.051 | 0.625 | 0.625 | 0.624 | 0.597 | 0.541 | 0.533 |
| B | 0.170 | 0.169 | 0.169 | 0.138 | 0.061 | 0.050 | 0.627 | 0.627 | 0.627 | 0.600 | 0.541 | 0.533 |
| C | 0.248 | 0.217 | 0.248 | 0.238 | 0.061 | 0.050 | 0.698 | 0.670 | 0.698 | 0.685 | 0.545 | 0.537 |
| D | 0.572 | -0.491 | 0.572 | 0.567 | 0.118 | 0.093 | 0.838 | 0.427 | 0.838 | 0.828 | 0.537 | 0.528 |
| E | 0.212 | 0.212 | 0.212 | 0.183 | 0.098 | 0.080 | 0.766 | 0.766 | 0.766 | 0.722 | 0.609 | 0.586 |
| F | 0.212 | 0.212 | 0.212 | 0.183 | 0.099 | 0.080 | 0.766 | 0.766 | 0.766 | 0.722 | 0.608 | 0.586 |
| G | 0.214 | 0.214 | 0.214 | 0.186 | 0.099 | 0.080 | 0.770 | 0.770 | 0.770 | 0.726 | 0.609 | 0.587 |
| H | 0.274 | 0.246 | 0.274 | 0.252 | 0.095 | 0.077 | 0.790 | 0.754 | 0.790 | 0.755 | 0.590 | 0.571 |

### 4.2.2 Simulation results

The histograms of $P$ and $P_h$ for the eight populations appear in Figure B.1 and Figure B.2. The impact of population HTE on $cfb$ and $C_b$ is summarized in Table 4.2. When there is no treatment interaction in the population ($\beta_{1,\Delta} = 0$), the larger the absolute main treatment effect ($|\beta_a|$), the greater the $C_b$ and $cfb$. As the value of $cfb$ increases as $|\beta_a|$ increases, the red curve shows a symmetric pattern. The value of $C_b$ increases as $\beta_a$ becomes more negative. However, $C_b$ turns to negative when the population-average benefit is negative. Figure 4.1 illustrates such phenomenon that refers to the part that the blue curve is below zero. This happens when the assumption made in Section 3.1.2 is violated. When $\beta_a > 0$, we have $E(B) \in (-0.6, -0.0995]$. Values of $C_b$ can be negative when a treatment does harm to the population on average, while some patients benefit from it. The $C_b$ aims to handle the scenarios that a new treatment therapy provides positive average benefit to the population. However, the treatment is only available to a subgroup of the population because of the high cost. Take COVID-19 vaccination as an example. Since it is impossible to vaccinate everyone simultaneously, we need to distribute the vaccines in phases. People who have high risk and are more vulnerable to severe illness are selected as priority targets.

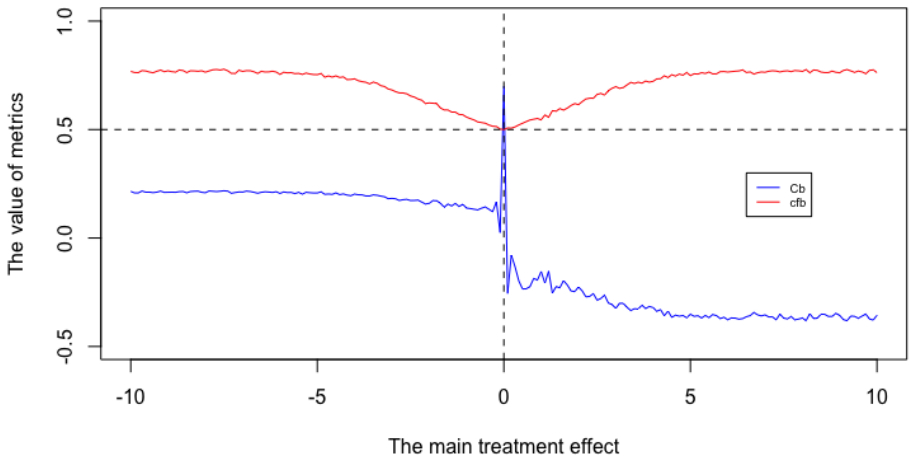



Figure 4.1: The values of $C_b$ and $cfb$ with different main treatment effects but no treatment interaction effect

Figure 4.2 illustrates several other curious features about the behaviour of $C_b$ and $cfb$. First, it shows upward tendencies of $C_b$ and $cfb$ when the absolute value of treatment interaction increases. Second, consider a specific value of the treatment interaction. If this interaction is fairly small (e.g. $\beta_{1,\Delta} = -0.05$), both $C_b$ and $cfb$ increase as $|\beta_a|$ increases. On the contrary, if this interaction term is large (e.g. $\beta_{1,\Delta} = -5$), the higher the $|\beta_a|$, the smaller the $C_b$ and $cfb$. Third, such small interaction ranges for $C_b$ and $cfb$ are different. Figure 4.2 shows that the range of $C_b$ is smaller than that of $cfb$ in this setup. It also explains the results refer to population $C$ and $G$ in Table 4.2: when $\beta_{1,\Delta} = -1$, the value of $C_b$ decreases from 0.248 to 0.214 as $|\beta_a|$ increases from 2 to 6. However, the value of $cfb$ increases from 0.698 to 0.770.

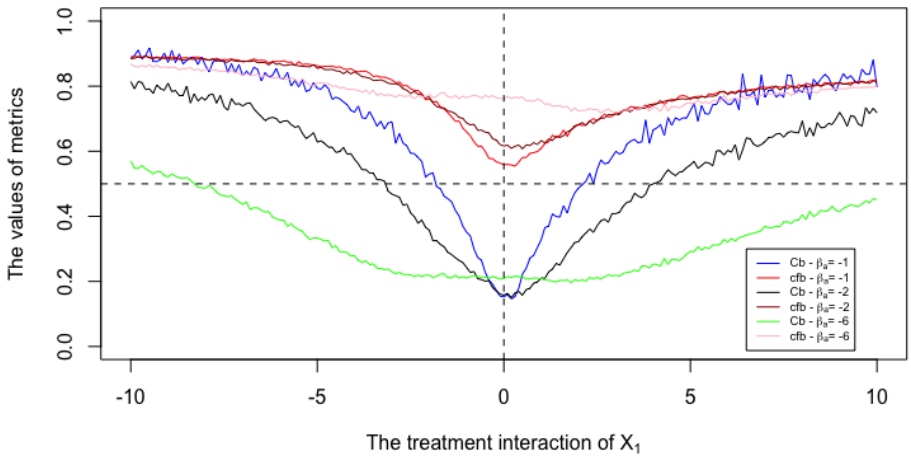



Figure 4.2: The values of $C_b$ and $cfb$ with different main treatment effects and treatment interaction effects

The impact of population HTE in $cfb_h, C_{b,h}$ also appears in Table 4.2. When there is no treatment interaction in the population, or when the interaction is negligible (i.e. population $A$, $B$, $E$, $F$ and $G$), the value of $C_{b,h}$ and $cfb_h$ are identical to $C_b$ and $cfb$ respectively, if $h$ describes the true relationship between response and covariates. In this setup, the first treatment benefit predictor that does not consider any treatment interaction performs well in explaining the underlying relationship of the data. However, we have $C_{b,1} < C_b$ and $cfb_1 < cfb$ when the treatment interaction cannot be negligible. If an interaction is surprisingly large, a benefit

predictor that fails to consider such interaction provides poor prediction performance. Whether the interaction is negligible or not is determined by $\beta_{1,\Delta}$ and $\beta_a$. When $\beta_a$ is large (e.g. $\beta_a = 6$), $\beta_{1,\Delta} = 1$ is counted as negligible interaction. Whereas, the $\beta_{1,\Delta} = 1$ cannot be ignored when $\beta_a$ is small (e.g. $\beta_a = 2$). Even if $E[B] > 0$, such poor benefit predictors may yield negative values of $cfb_h$ and $C_{b,h}$. For instance, in the population $D$ with the $\beta_a = 2$ and $\beta_{1,\Delta} = -4$, we compute $C_{b,1} = -0.491$ and $cfb_1 = 0.427$. To put it another way, when both values of $cfb_h$ and $C_{b,h}$ are falling outside of the defined boundaries, it implies that the relative benefit predictor might fail to describe the essential interaction effect.

Let's look at $cfb_h$ and $C_{b,h}$ for the other three benefit predictors in each population. These predictors relate to the regression models that contain one covariate and its corresponding treatment interaction term. The pattern that $C_{b,3} > C_{b,4} > C_{b,5}$ and $cfb_3 > cfb_4 > cfb_5$ appears in all eight populations. Such a pattern is determined by the covariate effect defined in the related fully specified model. Precisely, the benefit predictor that contains covariate with more information about $Y$ gives larger values of $C_{b,h}$ and $cfb_h$. We will see the similar results in Section 4.3.3.

Finally, we consider $\widehat{C}_{b,h}$ and $\widehat{cfb_h}$ of five benefit predictors in each population. The results for population A, B, C, and D appear in Table 4.3. We find that the standard deviation (SD) and the root mean square error (RMSE) for $\widehat{C}_{b,h}$ are getting larger when the benefit predictors become less informative. For $\widehat{cfb_h}$, values of SD are more stable over the five benefit predictors. The difference between SD and RMSE reflects the average bias of the predictor. Overall, $\widehat{cfb_h}$ provides slightly larger bias compared with $\widehat{C}_{b,h}$. The similar features of $\widehat{C}_{b,h}$ and $\widehat{cfb_h}$ can be found for the other four populations.

## 4.3 Simulation studies with correlation among counterfactual outcomes

### 4.3.1 Generate correlated counterfactual outcomes

To study the impact of correlated binary counterfactual outcomes on the three pair of metrics: $(cfb, C_b), (cfb_h, C_{b,h})$ and $(\widehat{cfb_h}, \widehat{C}_{b,h})$, an algorithm to generate a population with two correlated binary responses is needed. Since two counterfactual outcomes exist for the same individual under two treatment arms, it is rea-

Table 4.3: The values of $(C_{b,h}, cfb_h)$ and $(\widehat{C}_{b,h}, \widehat{cfb_h})$ for population A, B, C, and D

| | | $\widehat{C}_{b,h}$ | | | | | $\widehat{cfb_h}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $h$ | $C_{b,h}$ | $Median$ | $Mean$ | $SD$ | $RMSE$ | $cfb_h$ | $Median$ | $Mean$ | $SD$ | $RMSE$ |
| A | | | | | | | | | | |
| 1 | 0.167 | 0.162 | 0.163 | 0.051 | 0.051 | 0.625 | 0.605 | 0.606 | 0.034 | 0.038 |
| 2 | 0.167 | 0.166 | 0.165 | 0.050 | 0.050 | 0.624 | 0.607 | 0.607 | 0.034 | 0.038 |
| 3 | 0.135 | 0.129 | 0.131 | 0.053 | 0.053 | 0.597 | 0.583 | 0.584 | 0.035 | 0.037 |
| 4 | 0.062 | 0.066 | 0.060 | 0.060 | 0.060 | 0.541 | 0.534 | 0.534 | 0.035 | 0.035 |
| 5 | 0.051 | 0.048 | 0.041 | 0.066 | 0.066 | 0.533 | 0.524 | 0.524 | 0.034 | 0.035 |
| B | | | | | | | | | | |
| 1 | 0.169 | 0.169 | 0.167 | 0.051 | 0.051 | 0.627 | 0.610 | 0.611 | 0.035 | 0.039 |
| 2 | 0.169 | 0.165 | 0.167 | 0.049 | 0.049 | 0.627 | 0.606 | 0.609 | 0.033 | 0.038 |
| 3 | 0.138 | 0.143 | 0.141 | 0.054 | 0.054 | 0.600 | 0.592 | 0.591 | 0.038 | 0.039 |
| 4 | 0.061 | 0.059 | 0.058 | 0.062 | 0.062 | 0.541 | 0.535 | 0.535 | 0.034 | 0.034 |
| 5 | 0.050 | 0.047 | 0.040 | 0.065 | 0.066 | 0.533 | 0.521 | 0.522 | 0.034 | 0.036 |
| C | | | | | | | | | | |
| 1 | 0.231 | 0.227 | 0.226 | 0.038 | 0.039 | 0.671 | 0.657 | 0.660 | 0.031 | 0.033 |
| 2 | 0.298 | 0.291 | 0.292 | 0.040 | 0.041 | 0.732 | 0.714 | 0.713 | 0.027 | 0.034 |
| 3 | 0.290 | 0.287 | 0.287 | 0.041 | 0.041 | 0.722 | 0.703 | 0.702 | 0.028 | 0.034 |
| 4 | 0.061 | 0.068 | 0.063 | 0.054 | 0.054 | 0.544 | 0.543 | 0.540 | 0.034 | 0.034 |
| 5 | 0.050 | 0.052 | 0.046 | 0.057 | 0.057 | 0.535 | 0.531 | 0.530 | 0.033 | 0.033 |
| D | | | | | | | | | | |
| 1 | -0.491 | -0.498 | -0.630 | 0.534 | 0.551 | 0.427 | 0.412 | 0.413 | 0.032 | 0.035 |
| 2 | 0.572 | 0.569 | 0.577 | 0.062 | 0.063 | 0.838 | 0.821 | 0.822 | 0.022 | 0.027 |
| 3 | 0.567 | 0.553 | 0.558 | 0.055 | 0.056 | 0.828 | 0.813 | 0.814 | 0.020 | 0.025 |
| 4 | 0.118 | 0.121 | 0.112 | 0.099 | 0.099 | 0.537 | 0.533 | 0.536 | 0.033 | 0.033 |
| 5 | 0.093 | 0.086 | 0.080 | 0.109 | 0.110 | 0.528 | 0.529 | 0.528 | 0.033 | 0.032 |

sonable to consider the correlation between the two outcomes [27]. Therefore, we assume that the same individual's counterfactual outcomes are correlated, but the outcomes from different individuals are independent.

Table 4.4: Bivariate distribution for counterfactual outcomes

| $Y^{(0)}$ / $Y^{(1)}$ | 0 | 1 | |
|---|---|---|---|
| 0 | $a$ | $b$ | $1-p$ |
| 1 | $c$ | $d$ | $p$ |
| | $1-q$ | $q$ | 1 |

To generate the correlated counterfactual outcomes, we start with the bivariate binary distribution, which is shown in Table 4.4. In it, the marginal outcome probabilities are denoted as $Pr\{Y^{(0)}=1\}=p$ and $Pr\{Y^{(1)}=1\}=q$. The joint probabilities of counterfactual outcomes are denoted as a, b, c and d, respectively. Let $Pr\{Y^{(0)}=1, Y^{(1)}=1\}=d$. The boundary of $d$ for bivariate binary variables is known as the Fréchet-Hoeffding bounds [31], which takes form

$$\max\{0, p+q-1\} \leq d \leq \min\{p, q\}.$$

Based on the definition of the binary distribution, the means and variances of counterfactual outcomes can be calculated as:

$$E[Y^{(0)}] = p, Var(Y^{(0)}) = p(1-p),$$
$$E[Y^{(1)}] = q, Var(Y^{(1)}) = q(1-q).$$

The correlation coefficient of $Y^{(0)}$ and $Y^{(1)}$ is expressed as

$$r = \frac{Cov(Y^{(0)}, Y^{(1)})}{\sqrt{Var(Y^{(0)})}\sqrt{Var(Y^{(1)})}} = \frac{d-pq}{\sqrt{p(1-p)}\sqrt{q(1-q)}}. \tag{4.3}$$

Equation 4.3 shows that $r$ is not only determined by marginal probabilities $p$ and $q$, but it also depends on the joint probability $d$, which implies that the range of $r$ may be smaller than [-1,1]. Details of the boundary are given by [32], and the boundary can be expressed as:

$$\max\Big\{-M, -\frac{1}{M}\Big\} \leq r \leq -\sqrt{\frac{\min\{p,q\}(1-\max\{p,q\})}{\max\{p,q\}(1-\min\{p,q\})}},$$

where $M = \sqrt{\frac{pq}{(1-p)(1-q)}}$. Note that the boundary equals to $[-1, 1]$ when $p = q$.

Without considering the covariates information correlated bivariate binary counterfactual outcomes can be generated using the approach mentioned above. Several prognostic covariates are considered, and the proper relationship between the counterfactual outcomes and prognostic covariates is assumed logistic. The actual marginal probabilities for the $i$-th individual are defined as:

$$p_i = Pr\left\{Y_i^{(0)} = 1 | X_i, A_i = 0\right\} = \text{expit}\{X_i\beta\}, \tag{4.4}$$

$$q_i = Pr\left\{Y_i^{(1)} = 1 | X_i, A_i = 1\right\} = \text{expit}\{X_i\beta' + \beta_a\}, \tag{4.5}$$

where $X$ denotes a set of covariates, and the treatment assignment indicator is $A \in \{0, 1\}$. The coefficient $\beta$ reflects the effect of covariates on the binary outcome probabilities, $\beta'$ reflects the effect of the covariates under the treatment arm, and $\beta_a$ represents the main treatment effect.

Note that $Y^{(0)}$ and $Y^{(1)}$ generated from Equation 4.4 and Equation 4.5 are correlated as the probabilities are determined by the same covariates. Therefore, we focus on the correlation between outcomes that condition on the covariates. Let $\rho$ denote the level of conditional dependency. If $Y^{(0)}$ and $Y^{(1)}$ are conditionally independent, then $\rho = 0$.

The detail of data generation can be found in [33] [34] [35], and the brief summary of implementation steps used in this section are shown as follows:

Step 1. Per-determine population size $N$ and possible values of $\rho$, $\beta$, $\beta'$, and $\beta_a$. Three continuous standard normal distributed covariates, $X_1$, $X_2$, $X_3$, are generated.

Step 2. Two latent variables, $Z_0$ and $Z_1$, are generated from a bivariate standard normal distribution with correlation $\rho$. The bivariate normal distribution controls the level of dependence between the conditional counterfactual outcomes.

Step 3. Apply a transformation on $Z_0$ and $Z_1$ and get $\varepsilon_0 = \text{logit}(\Phi(Z_0))$ and $\varepsilon_1 = \text{logit}(\Phi(Z_1))$, where logit(.) is the inverse logistic function and $\Phi(.)$ is the cumulative distribution function.

Step 4. Calculate the log odds for counterfactual outcomes according to Equation 4.4 and Equation 4.5, and the log odds for $Y^{(0)}$ is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$,

and the log odds ratio for $Y^{(1)}$ is $\beta_0' + \beta_1' X_1 + \beta_2' X_2 + \beta_3' X_3 + \beta_a$.

Step 5. The two binary outcomes are then generated

$$Y^{(0)} = I(\varepsilon_0 \leq log(\frac{p}{1-p})),$$
$$Y^{(1)} = I(\varepsilon_1 \leq log(\frac{q}{1-q})),$$

where the function $I(.)$ is an indicator function.

### 4.3.2 Simulation setup

In the following simulation studies, all estimation results are based on independent 200 replicated data sets sampled from the population, and each data sets contains 500 subjects. The data set are sampled from a population with $10^5$ subjects using the data-generating mechanisms that are introduced in 4.3.1, and the population is described by

$$X_1 \sim N(0,1), X_2 \sim N(0,1), X_3 \sim N(0,1), \tag{4.6}$$
$$\text{logit}\left(E[Y^{(0)} = 1|X_1, X_2, X_3]\right) = X_1 + 0.5X_2 - 0.4X_3, \tag{4.7}$$
$$\text{logit}\left(E[Y^{(1)} = 1|X_1, X_2, X_3]\right) = X_1 + 0.5X_2 - 0.4X_3 + 1, \tag{4.8}$$

where $X_1, X_2, X_3$ are independent standard normal random variables, and $Y^{(0)}$ is correlated with $Y^{(1)}$ given the covariates. We consider four scenarios, and the correlation is set to be $\rho = 0, 0.3, 0.6, 0.9$, respectively. The correlation between $Z_0$ and $Z_1$, $\rho$, is controlled to roughly adjust the amount of the correlation between $Y^{(0)}$ and $Y^{(1)}$, denoted as $\delta$. The $\delta$ is less than $\rho$ due to the non-linear probability transformation applied to $Z_0$ and $Z_1$. We first calculate $C_b$ and $cfb$ for the population data in each scenario. Then we fit the five imposed models by fitting the imposed model structure to the population data and calculate corresponding $C_{b,h}$ and $cfb_h$. The last step is to estimate the $C_{b,h}$ and $cfb_h$ using the sample data sets. The simulation results for five imposed model structures are shown in Table 4.5.

Table 4.5: The values of $(C_b, cfb)$, $(C_{b,h}, cfb_h)$ and $(\widehat{C}_{b,h}, \widehat{cfb_h})$ for five imposed benefit predictors with different levels of correlation among counterfactual outcomes

| | | $\widehat{C}_{b,h}$ | | | | | $\widehat{cfb_h}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $h$ | $C_{b,h}$ | $Median$ | $Mean$ | $SD$ | $RMSE$ | $cfb_h$ | $Median$ | $Mean$ | $SD$ | $RMSE$ |
| *none* $C_b = 0.139$ and $cfb = 0.557$ | | | | | | | | | | |
| 1 | 0.139 | 0.132 | 0.135 | 0.106 | 0.106 | 0.557 | 0.540 | 0.543 | 0.029 | 0.032 |
| 2 | 0.139 | 0.143 | 0.140 | 0.125 | 0.125 | 0.562 | 0.543 | 0.544 | 0.033 | 0.036 |
| 3 | 0.100 | 0.088 | 0.081 | 0.116 | 0.116 | 0.542 | 0.527 | 0.526 | 0.030 | 0.032 |
| 4 | 0.042 | 0.037 | 0.031 | 0.131 | 0.132 | 0.516 | 0.511 | 0.510 | 0.029 | 0.030 |
| 5 | 0.030 | 0.038 | 0.010 | 0.212 | 0.212 | 0.512 | 0.512 | 0.511 | 0.032 | 0.032 |
| *weak* $C_b = 0.142$ and $cfb = 0.563$ | | | | | | | | | | |
| 1 | 0.142 | 0.154 | 0.147 | 0.116 | 0.115 | 0.562 | 0.544 | 0.546 | 0.032 | 0.036 |
| 2 | 0.142 | 0.150 | 0.143 | 0.103 | 0.103 | 0.562 | 0.548 | 0.548 | 0.029 | 0.033 |
| 3 | 0.100 | 0.108 | 0.102 | 0.147 | 0.148 | 0.542 | 0.533 | 0.534 | 0.034 | 0.035 |
| 4 | 0.042 | 0.044 | 0.024 | 0.147 | 0.148 | 0.516 | 0.514 | 0.513 | 0.030 | 0.030 |
| 5 | 0.030 | 0.032 | 0.029 | 0.135 | 0.135 | 0.512 | 0.513 | 0.514 | 0.031 | 0.031 |
| *moderate* $C_b = 0.141$ and $cfb = 0.572$ | | | | | | | | | | |
| 1 | 0.141 | 0.155 | 0.149 | 0.112 | 0.112 | 0.572 | 0.545 | 0.544 | 0.032 | 0.043 |
| 2 | 0.141 | 0.149 | 0.142 | 0.098 | 0.098 | 0.572 | 0.546 | 0.546 | 0.028 | 0.040 |
| 3 | 0.099 | 0.117 | 0.102 | 0.127 | 0.126 | 0.549 | 0.531 | 0.530 | 0.032 | 0.037 |
| 4 | 0.041 | 0.514 | 0.512 | 0.032 | 0.033 | 0.519 | 0.509 | 0.509 | 0.032 | 0.034 |
| 5 | 0.028 | 0.035 | 0.031 | 0.129 | 0.129 | 0.513 | 0.514 | 0.513 | 0.031 | 0.031 |
| *strong* $C_b = 0.140$ and $cfb = 0.594$ | | | | | | | | | | |
| 1 | 0.140 | 0.151 | 0.153 | 0.110 | 0.110 | 0.594 | 0.545 | 0.544 | 0.032 | 0.043 |
| 2 | 0.140 | 0.150 | 0.146 | 0.098 | 0.098 | 0.594 | 0.546 | 0.546 | 0.028 | 0.040 |
| 3 | 0.101 | 0.112 | 0.105 | 0.119 | 0.119 | 0.565 | 0.531 | 0.530 | 0.032 | 0.037 |
| 4 | 0.039 | 0.041 | 0.019 | 0.159 | 0.160 | 0.523 | 0.509 | 0.509 | 0.033 | 0.034 |
| 5 | 0.027 | 0.041 | 0.026 | 0.131 | 0.130 | 0.516 | 0.514 | 0.513 | 0.031 | 0.031 |

### 4.3.3 Simulation results

The simulation results can be found in Table 4.5. The most notable observation from the Table is that $cfb$ is influenced by the correlation between the $Y^{(0)}$ and $Y^{(1)}$, but $C_b$ is not. The value of $cfb$ increases when the correlation between counterfactual outcomes becomes larger, which increases from $0.563$ to $0.595$. Oppositely, the value of $C_b$ is around $0.140$ over the four correlation levels. The minor impact of the correlation on $C_b$ is due to simulation noise.

Take a detailed look at the $cfb_h$ and $C_{b,h}$ in Table 4.5. The benefit predictor and data determine the values of $cfb_h$ and $C_{b,h}$. When the benefit predictor describes the relationship between covariates and the treatment benefit, we have $cfb = cfb_h$ and $C_b = C_{b,h}$. Among all five benefit predictors in each scenario, we have $cfb \approx cfb_1 \approx cfb_2$ and $C_b \approx C_{b,1} \approx C_{b,2}$. The first predictor describes the actual association, and the second predictor is an enriched model of the first one with covariates and the first-order treatment interaction. However, when $h$ cannot describe the genuine relationship, we have $cfb > cfb_h$ and $C_b > C_{b,h}$. In our setup, The third, the fourth, and the fifth benefit predictors cannot predict the actual treatment benefit as they only contain a prognostic variable and its treatment interaction. Take the fourth benefit predictor as an example. We see $cfb > cfb_4$ and $C_b > C_{b,4}$. Note that histograms of $P$ and $P_h$ appear in Figure B.3.

Additionally, since all covariates are independently generated from the standard normal distribution, the values of $cfb_h$ and $C_{b,h}$ for $h = 3, 4, 5$ the extent of the covariates' effect on the response. The benefit predictor that contains more informative covariates has larger $cfb_h$ and $C_{b,h}$ values. In detail, the effects of $X_1$, $X_2$ and $X_3$ on the outcome are 1, 0.5, and -0.4, respectively. As a result, we have $cfb_3 > cfb_4 > cfb_5$ and $C_{b,3} > C_{b,4} > C_{b,5}$.

Now, let's look at the $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$. The SD of $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ over the four correlation levels are relatively stable. For $\widehat{C}_{b,h}$, the difference between SD and RMSE is negligible over all correlation levels. However, for $\widehat{cfb_h}$, RMSE and the difference between SD and RMSE increase as the outcomes correlation increases. In other words, the estimator of $cfb_h$ is sensitive to the correlation, and it provides biased estimates when there is a correlation between $Y^{(0)}$ and $Y^{(1)}$.

Another notable feature is that the value of SD and RMSE of the $\widehat{cfb_h}$ is smaller than that of $\widehat{C}_{b,h}$. This might be caused by several reasons. First, as the scales of $cfb$ and $C_b$ are different, a one unit (e.g. 0.01) increase in $cfb$ holds a different meaning, compared with the same amount of increase in $C_b$. The two metrics are

Table 4.6: The values of $C_b$ and $cfb$ with different levels of correlation among covariates

| *Corr/model* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Corr(X_1, X_2) = 0.8$ | | | | | |
| $C_b$ | 0.183 | 0.182 | 0.158 | 0.132 | 0.024 |
| $cfb$ | 0.575 | 0.574 | 0.563 | 0.551 | 0.508 |
| $Corr(X_1, X_2) = -0.8$ | | | | | |
| $C_b$ | 0.086 | 0.087 | 0.060 | 0.029 | 0.047 |
| $cfb$ | 0.535 | 0.535 | 0.524 | 0.511 | 0.518 |
| $Corr(X_1, X_3) = 0.8$ | | | | | |
| $C_b$ | 0.101 | 0.102 | 0.073 | 0.0453 | 0.042 |
| $cfb$ | 0.541 | 0.541 | 0.528 | 0.517 | 0.515 |

defined in different fixed ranges, where $cfb \in [0.5, 1]$ and $C_b \in [0, 1]$. Second, for $C_{b,h}$, we assume $E[B] > 0$. This assumption might be violated for some samples, which leads to corresponding estimates falling outside of the boundary. All in all, the SD values of $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ are not comparable.

**What if $X_1$, $X_2$, and $X_3$ are not pairwise independent?**

We have explored the impact of counterfactual outcome correlation on $(cfb, C_b)$, $(cfb_h, C_{b,h})$ and $(\widehat{cfb_h}, \widehat{C}_{b,h})$, when there is no correlation between any covariate pairs. What if some corresponding covariates are correlated? To simplify the problem, we assume there is no correlation between counterfactual outcomes, and let $corr_x$ denote the pairwise correlation amongst covariates. First, we consider three different levels of correlation, which are $corr(X_1, X_2) = 0, 0.3, 0.8$, respectively. The results appear in Table B.3. Second, we compare the metrics when $corr(X_1, X_2) = 0.8$ with the metrics when $corr(X_1, X_2) = -0.8$ and $corr(X_1, X_3) = 0.8$, respectively. The results of $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ are shown in Table 4.6.

Table B.3 shows that $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ increase as $corr(X_1, X_2)$ increases from 0 to 0.8 when $h = 1, 2, 3, 4$. When $h = 5$, we found decreases in values of $C_{b,h}$ and $cfb_h$. Similarly, the estimator of $C_{b,h}$ is not influenced by $corr_x$, and the SD of $\widehat{C}_{b,h}$ is close to its RMSE. However, for $\widehat{cfb_h}$, the difference between SD and RMSE increases when $corr(X_1, X_2)$ increases. Recall that we fit

the population data to the imposed model structures to obtain the imposed models. When covariates are highly correlated, the estimated model coefficients are different from the actual effects, which does not reflect the genuine relationship between the corresponding covariate and the response.

The differences between $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ under these three cases:

$$corr(X_1, X_2) = -0.8, \;\; corr(X_1, X_3) = 0.8, \;\; \text{and } corr(X_1, X_2) = 0.8,$$

appear in Table 4.6. We can see that $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ decrease when the correlation change from $corr(X_1, X_2) = 0.8$ to $corr(X_1, X_2) = -0.8$ or to $corr(X_1, X_3) = 0.8$. The results are reasonable as $X_1$ and $X_2$ have positive effects on the response. If $X_1$ and $X_2$ are positively correlated, the values of $(cfb, C_b)$ and $(cfb_h, C_{b,h})$ increase. Instead, if they are negatively highly correlated, the corresponding values decrease. Similarly, since $X_3$ has negative effects on the response, metrics values decrease if $X_1$ and $X_3$ are positively correlated. To sum up, the correlation between covariates affects $(cfb, C_b)$ and $(cfb_h, C_{b,h})$, and could lead to an less precise estimation.

# Chapter 5

# Case study: analysis of the acute myocardial infarction data

We apply $\widehat{cfb}_h$ and $\widehat{C}_{b,h}$ to the acute myocardial infarction data to illustrate how these two metrics work in practice. Acute myocardial infarction (MI) is one of the common causes of death, which occurs from an acute cut-off of the heart muscle's blood supply. The data are from the GUSTO-I study, which aimed to investigate and compare the relative efficacy of several treatment interventions for acute MI [14]. This was a randomized controlled trial with data from 1081 hospital sites in 15 countries. The primary response variable is patient's death status (1 = death, 0 = alive), measured at the 30-th day after the onset of the disease. The data set, containing 40,830 patients and 2851 deaths, has been widely used to study different aspects of regression modelling, such as model internal validation [36].

This analysis focuses on two interventions: the old standard (streptokinase and intravenous heparin) and the primary new treatment (accelerated t-PA and intravenous heparin). As a result, the large data set narrows down to 30510 patients with 2128 deaths. Additionally, 27 clinical measurements were recorded, including age, family history, smoking status, and hypertension.

To demonstrate $\widehat{cfb}_h$ and $\widehat{C}_{b,h}$ for an externally defined benefit predictor, these 30510 patients were randomly split into a training set and test set. The training data can be thought of as a previous trial, where several treatment benefit predictors have been built based on patients' clinical measurements. The test data can be thought of as another similar trial. We evaluate the discriminatory ability of these

benefit predictors obtained from the previous trial on the test data by calculating the metrics of interest. There are 16000 patients in the training data made up of 8000 patients randomly selected from the SK treatment arm and the other 8000 patients randomly chosen from the tPA treatment arm. The test data contains the 14510 patients who were not in the training data. In the training data, the estimated $E(B)$ is

$$\begin{aligned}\widehat{E}(B)_{\text{train}} =& \frac{\sum_{i=1}^{16000} Y_i I(A_i = 0)}{\sum_{i=1}^{16000} I(A_i = 0)} - \frac{\sum_{i=1}^{16000} Y_i I(A_i = 1)}{\sum_{i=1}^{16000} I(A_i = 1)} \\ &= 0.0732 - 0.0635 \\ &= 0.0097.\end{aligned}$$

And in the test data, we have

$$\begin{aligned}\widehat{E}(B)_{\text{test}} =& \frac{\sum_{j=1}^{14510} Y_j I(A_j = 0)}{\sum_{j=1}^{14510} I(A_j = 0)} - \frac{\sum_{j=1}^{14510} Y_j I(A_j = 1)}{\sum_{j=1}^{14510} I(A_j = 1)} \\ &= 0.0731 - 0.0618 \\ &= 0.0113.\end{aligned}$$

The benefit predictors are developed based on fully specified multivariable logistic regression models. To perform a multivariable regression, we may consider whether all clinical measurements and related pair-wise treatment interactions are associated with the 30-day mortality. Covariates are denoted as $X_1, X_2, ..., X_m$ with $m = 27$. To describe the association between the death and a set of the covariates, for the $i$-th individual, we start by the model in the form of

$$\text{logit}(E(Y_i|X_i, A_i)) = \beta_0 + \beta_a A_i + \sum_{j=1}^{m} \beta_j X_i + \sum_{j=1}^{m} \beta_{a,j} X_i A_i, \tag{5.1}$$

where $\beta_0$ is the intercept; $\beta_a$ is the main treatment effect; $\beta_j$ is the main covariates effect; and $\beta_{a,j}$ is the first-order treatment interaction effect.

The model is improved by using a subset of the clinical measurements and interactions (predictors) significantly related to death. There are various types of variable selection procedures. One widely used method is the stepwise variable selection using criteria that assess the goodness-of-fit of a model and the model complexity. Here, we consider both the stepwise variable selection using Akaike Information Criterion (AIC) and the variable selection based on p-values of Wald tests and

bootstrapping.

The Wald test is conducted for each predictor. If the p-value of an individual's coefficient estimator is below 0.05, the corresponding covariate effect on the 30-day mortality is significant. However, such significance might be due to chance, especially when the number of predictors is large. Therefore, we perform bootstrap sampling to avoid such instability. We fit the generalized linear model to 250 bootstrap samples and check the probability of p-value below 0.05 for each predictor. The predictors that satisfy $P(\text{p-value} < 0.05) \geq 0.8$ are considered.

The selected eight covariates are age, time to relief of chest pain ($>$ 1hr), systolic blood pressure (mmHg), shocking, heart rate (beats/min), hypertension, Killip class, and previous MI. Although no corresponding interaction terms are selected by the variable selection procedure, we consider three models:

- Model 1 contains these eight covariates.
- Model 2 contains the eight covariates and the treatment interactions with Killip class.
- Model 3 contains the eight covariates and all first-order treatment interactions.

Since these models are getting complex, especially Model 3, we intend to constrain the coefficient estimates to avoid the potential overfitting problem. Hence, model parameters are estimated by the standard unconstrained maximum likelihood and the ridge regression with $l_2$ norm shrinkage penalty. For the ridge regression, the value of the tuning parameter is selected by 10-fold cross-validation. Table 5.1 provides the estimates of regression coefficients. As cross-validation provides random folds in ridge regression models, this randomness leads to various results. Hence, the tuning parameter is calculated as an average of 25 replicates.

Six columns in Table 5.1 correspond to six different benefit predictors denoted as $h = ml_1, ml_2, ml_3, r_1, r_2, r_3$. The $ml_1, ml_2, ml_3$ are related to the three logistic models, and the $r_1, r_2, r_3$ are related to the three ridge regression models. We calculated the predicted treatment benefits as functions of covariates from the benefit predictors. Figure 5.1 demonstrates histograms of predicted treatment benefit $P_h$ from the six benefit predictors. In each histogram, we compared $P_h$ calculated using the training data with that calculated using the test data. The $P_h$ calculated using the test data is in red, and the other is in dark blue.

Table 5.1: Regression coefficients for the ridge and maximum likelihood estimates

| | *Unconstrained ML* | | | *Ridge regression* | | |
|---|---|---|---|---|---|---|
| *Variables \Model* | 1 | 2 | 3 | 1 | 2 | 3 |
| Intercept | -3.784 | -3.774 | -3.662 | -3.676 | -3.684 | -3.641 |
| Main effects | | | | | | |
| tx(tPA) | -0.141 | -0.159 | -0.411 | -0.196 | -0.175 | -0.264 |
| age | 0.947 | 0.947 | 0.864 | 0.972 | 0.973 | 0.813 |
| sho(yes) | 2.044 | 1.620 | 1.693 | 0.822 | 0.800 | 0.820 |
| pulse | 0.398 | 0.397 | 0.394 | 0.357 | 0.356 | 0.303 |
| pmi(yes) | 0.420 | 0.418 | 0.447 | 0.391 | 0.394 | 0.394 |
| hyp (yes) | 0.581 | 0.581 | 0.504 | 0.699 | 0.699 | 0.828 |
| Killip(II) | 0.603 | 0.601 | 0.614 | 0.611 | 0.663 | 0.680 |
| Killip(III) | -0.659 | -0.184 | -0.243 | 0.419 | 0.514 | 0.537 |
| Killip(IV) | | | | 1.131 | 0.831 | 0.842 |
| sysbp | -0.277 | -0.275 | -0.249 | -0.227 | -0.227 | -0.174 |
| ttr | 0.563 | 0.560 | 0.492 | 0.486 | 0.482 | 0.433 |
| Interactions | | | | | | |
| age | | | 0.183 | | | 0.034 |
| sho(yes) | | | | | | 0.000 |
| pulse | | | 0.016 | | | 0.116 |
| pmi(yes) | | | -0.125 | | | 0.000 |
| hyp | | | -0.047 | | | -0.283 |
| Killip(II) | | 0.004 | -0.029 | | -0.114 | -0.158 |
| Killip(III) | | -0.100 | -0.121 | | -0.151 | -0.254 |
| Killip(IV) | | 0.798 | 0.653 | | 0.564 | 0.475 |
| sysbp | | | 0.050 | | | -0.119 |
| ttr | | | -0.082 | | | 0.099 |

[a] All continuous variables are scaled with zero mean and unit standard deviation before fitting a model.

[b] We use the 'glmnet' function in the glmnet package to constrain the coefficients.

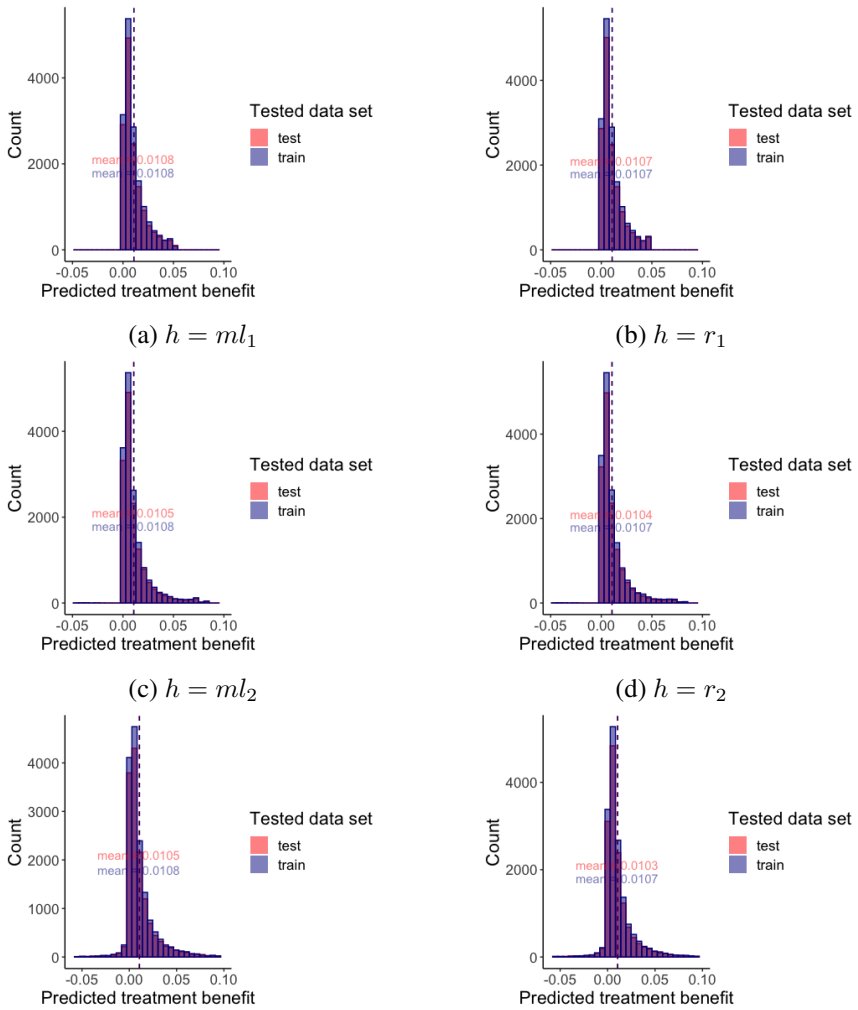
[c] Meanings of the selected eight covariates are as follows: age, shocking, heart rate (beats/min), previous MI, hypertension, Killip class, systolic blood pressure (mmHg), and time to relief of chest pain $> 1$ hr.

Table 5.2: The values of $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ for six imposed benefit predictors on the training and test data

| | $ml_1$ | $ml_2$ | $ml_3$ | $r_1$ | $r_2$ | $r_3$ |
|---|---|---|---|---|---|---|
| *training* | | | | | | |
| $\widehat{cfb_h}$ | 0.5198 | 0.5292 | 0.5287 | 0.5200 | 0.5278 | 0.5284 |
| *sd* | 0.0211 | 0.0211 | 0.0246 | 0.0250 | 0.0246 | 0.0235 |
| $\widehat{C}_{b,h}$ | 0.2941 | 0.4173 | 0.4441 | 0.2941 | 0.4171 | 0.4378 |
| *sd* | 0.0922 | 0.3381 | 0.5682 | 0.2412 | 0.1707 | 0.4226 |
| *test* | | | | | | |
| $\widehat{cfb_h}$ | 0.5227 | 0.5266 | 0.5103 | 0.5121 | 0.5313 | 0.4906 |
| *sd* | 0.0253 | 0.0171 | 0.0228 | 0.0262 | 0.0216 | 0.0278 |
| $\widehat{C}_{b,h}$ | 0.1190 | 0.1409 | | 0.1180 | 0.1410 | |
| *sd* | 0.2578 | 0.4907 | | 0.1379 | 0.2217 | |

[a] Some cells are empty as the numerical instability of $\widehat{C}_{b,h}$ becomes more pronounced as the model becomes more complex.

We estimated $cfb_h$ and $C_{b,h}$ for the imposed six benefit predictors using the training and test data, with results are shown in Table 5.2. We obtain the standard deviations of the estimators using the bootstrapping approach. It is reasonable that $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ for the training data are generally larger than the corresponding values of the test data. The numerical instability of $\widehat{C}_{b,h}$ becomes more pronounced as the model becomes more complex. Further insights into this instability will be sought in the future. The most complex model (Model 3) has the largest $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$ for the training data. However, the results of test data metrics imply that this model might overfit the data. It seems that Model 2 provides the best discriminatory ability among the three.




(a) $h = ml_1$ (b) $h = r_1$

(c) $h = ml_2$ (d) $h = r_2$

(e) $h = ml_3$ (f) $h = r_3$

Figure 5.1: Histograms of predicted treatment benefit $(P_h)$ calculated based on training and test data

# Chapter 6

# Conclusion and discussion

We have expanded the initial definitions of concentration of benefit index in [12] and concordance-statistic for benefit in [11] to three pairs of related metrics: $(cfb, C_b)$, $(cfb_h, C_{b,h})$, and $(\widehat{cfb_h}, \widehat{C}_{b,h})$. The first pair of metrics, $(cfb, C_b)$, is the ultimate goal if we are interested in answering questions such as: what is the association between the conditional expectation of treatment benefit given covariates $(E[B|X])$ and the actual treatment benefit $(B)$? Or whether the heterogeneity of $E[B|X]$ reflects the heterogeneity of $B$? As the model-based approach is the most common method to estimate $E[B|X]$, we introduced the second pair of metrics that refer to the parametric treatment benefit predictor, $(cfb_h, C_{b,h})$. The second pair of metrics depends on the benefit predictor $(h)$ which is a function of covariates. They summarize the association between the predicted treatment benefit $P_h$ and $B$ and assess the benefit predictor's predictive performance. The last pair of metrics, $(\widehat{cfb_h}, \widehat{C}_{b,h})$, are the estimators of $(cfb_h, C_{b,h})$, which are calculated using sample observations. The initial concordance-statistic for benefit in [11] refers to $\widehat{cfb_h}$, and the initial concentration of benefit index in [12] refers to $C_{b,h}$ and $\widehat{C}_{b,h}$.

These three pairs of metrics were explored conceptually (Chapter 2 and 3) and empirically (Chapter 4). The $cfb$ is defined based on the c statistic, and we studied the properties of c statistic in Section 2.2.1. Metrics related to $cfb$ play a similar rule as the c statistic, and they consider every patient's benefits with a noise which is the difference between $P$ and $B$. For a patient, if the noise is large, the medicine may not benefit the patient. In contrast, $C_b$ is averaging across the noise. As the Gini-like index, $C_b$ and its related metrics are of service to assess decision policies of using some medicines. For instance, the larger value of $C_b$ implies a stronger

justifications for personalized medicine, to maximize the population-level average treatment benefit with limited medical resources.

In two simulation studies of Chapter 4, we demonstrated the impact of heterogeneity of treatment effect and the counterfactual outcomes correlation on $(cfb, C_b)$, $(cfb_h, C_{b,h})$, and $(\widehat{cfb_h}, \widehat{C}_{b,h})$. We found that $(cfb, C_b)$ increase as the absolute main treatment effect increases when we fixed the treatment interaction term to a small value. When the interaction is very large, $cfb$ and $C_b$ decrease as the absolute main treatment effect increases. According to patients' covariates, the values of $cfb_h$ and $C_{b,h}$ reach up to $cfb$ and $C_b$ if the $h$ fully reflects the actual benefit. The benefit predictor performance also exhibits how well covariates can help predict the treatment benefit. The benefit predictor with more informative covariates has large $cfb_h$ and $C_{b,h}$. The standard deviation of $\widehat{C}_{b,h}$ is generally larger than that of $\widehat{cfb_h}$ because $C_b$ related metrics are ratio statistics. The denominator of ratio statistic can approach zero, resulting in large values in some bootstrap samples. The estimators, $\widehat{cfb_h}$ and $\widehat{C}_{b,h}$, are unbiased if there is no apparent correlation between counterfactual outcomes or covariates. Otherwise, the $cfb$ will be influenced by the correlation between counterfactual outcomes. Moreover, the estimator of $cfb_h$ is also sensitive to the correlation, and it may be biased. On the other hand, $C_{b,h}$ is unaffected by the correlation between counterfactual outcomes.

The method proposed in this thesis can be naturally extended to other types of response, such as count or time-to-event response. The corresponding benefit predictors can then be constructed by generalized linear models or Cox proportional hazards regression models.

This thesis's scope is kept manageable by only considering the RCT data, which guarantees that an association difference can identify the treatment benefit. It is not always easy to randomize the treatment for all problems. Therefore, the next step is to expand the scope by considering observational studies, where association may not imply causation. The apparent association might be due to confounding effect or some mixture of confounding and causation. To study the causal quantity, treatment benefit, we need to contemplate methods for controlling the observed and unobserved confounding effect interference [37]. Reconciling the methods for controlling confounding with the algorithms for estimating $cfb_h$ and $C_{b,h}$ remains an important topic to explore.

According to the definition, $\widehat{cfb_h}$'s sensitivity to the matching procedures should be carefully evaluated in future studies. To estimates the actual treatment benefit, we consider a matching procedure where the distance between two patients is

quantified by the difference of their predicted treatment benefits. We then use this distance to match patients using the nearest distance matching algorithm. We also consider matching patients by the probability of receiving the treatment accounting for the predicted treatment benefit, which yields slightly better results than that matched by the predicted treatment benefit. Take the population A and C in Section 4.2 as an example. We evaluate the behaviour of the two matching methods, and the results appear in Table 6.1. The implications of using other matching algorithms can be explored in future studies. We could consider diverse approaches to quantify the similarity of two patients or apply different matching algorithms [38].

Other research on this topic can focus on how to construct a benefit predictor when the interest is in evaluating the capacity of covariates in explaining HTE. There are three possibilities for developing a benefit predictor, and we have discussed the first one in this thesis. That is, when a benefit predictor is a known function either from previous experience or from other external knowledge on the subject, we then use estimates of $cfb_h$ and $C_{b,h}$ to assess the predictive performance of this known function. The function could refer to a fully specified model with known coefficients and model structure, which relates to what we have done in Chapter 4 and 5. Second, if we only know which covariates are in the set of predictors, then the coefficients of a benefit predictor and $cfb_h$ (or $C_{b,h}$) are estimated simultaneously. Such a process is applied in the simulation studies in [12]. Third, if the external knowledge provides a set of covariates with both predictors and irrelevant variables, the unnecessary variables should be excluded from the benefit predictor. By reducing the unnecessary complexity, we can obtain a benefit predictor with a more accurate prediction performance. In this case, the coefficients of a benefit predictor and $cfb_h$ (or $C_{b,h}$) are still estimated simultaneously, but some coefficients are allowed to be zero. LASSO is a method that selects variables by shrinking some coefficients to zero. For personalized medicine studies, interventions are the primary interests. There is some concern, however, that LASSO may provide a model without the intervention of interest, simply because all parameters are equally likely to be penalized to select the informative variables. Thus, it is reasonable to consider some LASSO related methods (e.g. group LASSO) to provide models that always contain the intervention [39]. One challenge with conventional LASSO is that for a model that includes both the main treatment effect and covariate-by-treatment interactions, the main effect might be removed. The hierarchy of covariates in the model might be violated, which we should wary of model interpretation. There could be more thinking about whether interaction and main effects should be grouped in LASSO.

Table 6.1: The values of $cfb_h$ and $\widehat{cfb_h}$ for population A and C using different matching methods

| | | *Modified version*[a] | | | | *Predicted treatment benefit* | | | |
|---|---|---|---|---|---|---|---|---|---|
| $h$ | $cfb_h$ | $Median$ | $Mean$ | $SD$ | $RMSE$ | $Median$ | $Mean$ | $SD$ | $RMSE$ |
| A | | | | | | | | | |
| 1 | 0.167 | 0.609 | 0.611 | 0.033 | 0.035 | 0.605 | 0.606 | 0.034 | 0.038 |
| 2 | 0.167 | 0.612 | 0.612 | 0.033 | 0.036 | 0.607 | 0.607 | 0.034 | 0.038 |
| 3 | 0.135 | 0.589 | 0.589 | 0.034 | 0.035 | 0.583 | 0.584 | 0.035 | 0.037 |
| 4 | 0.062 | 0.539 | 0.540 | 0.033 | 0.033 | 0.534 | 0.534 | 0.035 | 0.035 |
| 5 | 0.051 | 0.528 | 0.529 | 0.035 | 0.035 | 0.524 | 0.524 | 0.034 | 0.035 |
| C | | | | | | | | | |
| 1 | 0.671 | 0.667 | 0.667 | 0.032 | 0.032 | 0.657 | 0.660 | 0.031 | 0.033 |
| 2 | 0.732 | 0.718 | 0.717 | 0.028 | 0.032 | 0.714 | 0.713 | 0.027 | 0.034 |
| 3 | 0.722 | 0.705 | 0.707 | 0.029 | 0.033 | 0.703 | 0.702 | 0.028 | 0.034 |
| 4 | 0.544 | 0.545 | 0.546 | 0.034 | 0.034 | 0.543 | 0.540 | 0.034 | 0.034 |
| 5 | 0.535 | 0.537 | 0.534 | 0.033 | 0.033 | 0.531 | 0.530 | 0.033 | 0.033 |

[a] The modified version refers to the probability of receiving the treatment accounting for the predicted treatment benefit.

# Bibliography


[1] Jean-charles Sanchez, Matthias Doering, and Till Multiclass. *pROC: Title Display and Analyze ROC Curves*, 2021. R package version 1.17.0.1. → pages ix, 7

[2] Zeileis Achim and Christian Kleiber. *ineq: Measuring Inequality, Concentration, and Poverty*, 2014. R package version 0.2-13. → pages ix, 9

[3] Alexander Pate, Richard Emsley, Darren M. Ashcroft, Benjamin Brown, and Tjeerd Van Staa. The uncertainty with using risk prediction models for individual decision making: An exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Medicine*, 17(1):134, 2019. → page 1

[4] Menelaos Pavlou, Gareth Ambler, Shaun R Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z Omar. How to develop a more accurate risk prediction model when there are few events. *BMJ*, 351, 2015. → page 1

[5] Isabelle Kaiser, Annette B. Pfahlberg, Wolfgang Uter, Markus V. Heppt, Marit B. Veierød, and Olaf Gefeller. Risk prediction models for melanoma: A systematic review on the heterogeneity in model development and validation. *International Journal of Environmental Research and Public Health*, 17(21):1–25, nov 2020. → page 1

[6] Richard D Riley, Kym IE Snell, Karel GM Moons, and Thomas PA Debray. Fundamental statistical methods for prognosis research. In *Prognosis Research in Health Care*, chapter 3, pages 37–68. Oxford University Press, 2019. → pages 1, 5

[7] 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 traditional and novel measures. *Epidemiology*, 21(1):128–138, jan 2010. → page 1

[8] Ravi Varadhan and John D Seeger. Estimation and reporting of heterogeneity of treatment effects. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*, chapter 3, pages 35–44. Agency for Healthcare Research and Quality (US), 2013. → page 2

[9] Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and L. J. Wei. Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539, jun 2013. → page 2

[10] David M. Kent, Ewout Steyerberg, and David Van Klaveren. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ*, 363, dec 2018. → pages 2, 15

[11] David van Klaveren, Ewout W. Steyerberg, Patrick W. Serruys, and David M. Kent. The proposed a 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *Journal of Clinical Epidemiology*, 94:59–68, feb 2018. → pages 2, 7, 14, 15, 27, 33, 54

[12] Mohsen Sadatsafavi, Mohammad Ali Mansournia, and Paul Gustafson. A threshold-free summary index for quantifying the capacity of covariates to yield efficient treatment rules. *Statistics in Medicine*, 2020. → pages 2, 13, 15, 18, 19, 29, 54, 56, 67

[13] Lidia Ceriani and Paolo Verme. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, 10(3):421–443, sep 2012. → pages 3, 8

[14] The GUSTO Investigators. An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction. *New England Journal of Medicine*, 329(10):673–682, sep 1993. → pages 3, 48

[15] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr, and Douglas G. Manuel. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, 20(1):4, jan 2020. → page 6

[16] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005. → page 7

[17] Zhezhen Jin and Mounir Mesbah. Unidimensionality, agreement and concordance probability. In *Statistical Models and Methods for Reliability and Survival Analysis*, chapter 1, pages 1–19. Wiley Online Library, 2013. → page 7

[18] Matthias Schmid, Marvin N. Wright, and Andreas Ziegler. On the use of Harrell's C for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450–459, jul 2016. → page 7

[19] Carles M. Cuadras, losep Fortiana, and lose A. Rodriguez-lallena. *Distributions With Given Marginals and Statistical Modelling*. Springer Netherlands, 2002. → page 7

[20] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. CRC press, 2020. → page 7

[21] Roger B. Newson. Comparing the predictive powers of survival models using Harrell's C or Somers' D. *Stata Journal*, 10(3):339–358, sep 2010. → page 7

[22] Joseph L Gastwirth. A general definition of the lorenz curve. *Econometrica*, 39(6):1037–39, 1971. → page 8

[23] Kuan Xu. How has the literature on gini's index evolved in the past 80 years? *Dalhousie University, Economics Working Paper*, 2003. → pages 9, 10

[24] Shlomo Yitzhaki and Ingram Olkin. Concentration indices and concentration curves. *Lecture Notes-Monograph Series*, 19:380–392, 1991. → page 10

[25] Somesh Das Gupta. Gini association and pseudo lorenz curve. *Communications in Statistics - Theory and Methods*, 28(9):2181–2199, 1999. → page 10

[26] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman  Hall/CRC, 2020. → page 11

[27] Sander Greenland, Michael P. Fay, Erica H. Brittain, Joanna H. Shih, Dean A. Follmann, Erin E. Gabriel, and James M. Robins. On Causal Inferences for Personalized Medicine: How Hidden Causal Assumptions Led to Erroneous Causal Claims About the D-Value. *The American Statistician*, 74(3):243–248, 2020. → pages 15, 41

[28] Raaz Dwivedi, Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, and Bin Yu. Stable discovery of interpretable subgroups via

calibration in causal studies. *International Statistical Review*, 88(S1):S135–S178, 2020. → page 30

[29] Frank E. Harrell. *Hmisc: Harrell Miscellaneous*, 2021. R package version 4.5-0. → page 32

[30] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference*, 2020. R package version 4.1.0. → page 33

[31] Roger B. Nelsen, José Juan Quesada-Molina, José Antonio Rodríguez-Lallena, and Manuel Úbeda-Flores. Bounds on bivariate distribution functions with given margins and measures of association. *Communications in Statistics - Theory and Methods*, 30(6):1055–1062, 2001. → page 41

[32] Harry Joe. *Multivariate Models and Dependence Concepts*, volume 93. CRC Press, 1997. → page 41

[33] Lawrence J Emrich and Marion R Piedmonte. A method for generating high-dimensional multivariaten binary variates. *American Statistician*, 45(4):302–304, 1991. → page 42

[34] Glen Philip Martin, Matthew Sperrin, Kym I.E. Snell, Iain Buchan, and Richard D Riley. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Statistics in Medicine*, 2020. → page 42

[35] Anestis Touloumis. Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. *R Journal*, 8(2):79–91, 2016. → page 42

[36] Ewout W. Steyerberg, Frank E. Harrell, Gerard J.J.M. Borsboom, M. J.C. Eijkemans, Yvonne Vergouwe, and J. Dik F. Habbema. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–781, 2001. → page 48

[37] Zhi Geng, Yue Liu, Chunchen Liu, and Wang Miao. Evaluation of causal effects and local structure learning of causal networks. *Annual Review of Statistics and Its Application*, 6(1):103–124, 2019. → page 55

[38] Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, feb 2010. → page 56

[39] Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19(1):322, dec 2018. → page 56

[40] John Kloke and Joseph W. McKean. *Nonparametric Statistical Methods Using R*. CRC Press, 2014. → page 65

[41] Saralees Nadarajah and Samuel Kotz. Exact Distribution of the Max/Min of Two Gaussian Random Variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(2):210–212, 2008. → page 70

# Appendix A

# Mathematical details

## A.1 Connection between the c statistic and the Gini-like index

The c statistic and the Gini-like index are connected with each other when what we want to discriminate is binary. Consider two-subject experiments with a binary outcome $Y$ and $E[Y] = \pi$. Each two-subject pair contains $\{(Y_1, R_1), (Y_2, R_2)\}$, where $R_1 = E[Y_1|X_1]$ and $R_2 = E[Y_2|X_2]$. According to Equation 2.2, the c statistic for $R$ and $Y$ is:

$$\begin{aligned} c &= Pr\{R_2 > R_1 | Y_1 = 0, Y_2 = 1\} \\ &= \frac{Pr\{Y_1 = 0, Y_2 = 1 | R_2 > R_1\} Pr\{R_2 > R_1\}}{Pr\{Y_1 = 0, Y_2 = 1\}} \\ &= \frac{Pr\{Y_1 = 0, Y_2 = 1 | R_2 > R_1\}}{2\pi(1-\pi)}, \end{aligned}$$

and Gini-like index is expressed as

$$\begin{aligned} \text{Gini-like index} &= \frac{E[Y_2 - Y_1 | R_2 > R_1]}{2E[Y]} \\ &= \frac{Pr\{Y_1 = 0, Y_2 = 1 | R_2 > R_1\}}{2E[Y]} - \frac{Pr\{Y_1 = 1, Y_2 = 0 | R_2 > R_1\}}{2E[Y]}. \end{aligned}$$

Notice that these two expressions have the same term $Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}$, and the Gini-like index has an additional term $Pr\{Y_1 = 1, Y_2 = 0|R_2 > R_1\}$. To further explore the connection between the c statistic and the Gini-like index, we need to find the connection between $Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}$ and $Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}$. To begin with, as $Y_1$ and $Y_2$ are *iid*, we have:

$$Pr\{Y_1 = 0, Y_2 = 0|R_2 > R_1\} = Pr\{Y_1 = 0, Y_2 = 0|R_2 < R_1\}.$$

Moreover, it is reasonable to assume no ties in risk scores ($R_2 \neq R_1$), since $R$ is continuous. Then, we have

$$\begin{aligned}
&Pr\{Y_1 = 0, Y_2 = 0|R_2 > R_1\} \\
&= 2Pr\{Y_1 = 0, Y_2 = 0|R_2 > R_1\} \\
&= Pr\{Y_1 = 0, Y_2 = 0\} \\
&= Pr\{Y_1 = 0\}Pr\{Y_2 = 0\} \\
&= (1 - \pi)^2.
\end{aligned}$$

Similarly, $Pr\{Y_1 = 1, Y_2 = 1|R_2 > R_1\}$ is expressed as:

$$\begin{aligned}
&Pr\{Y_1 = 1, Y_2 = 1|R_2 > R_1\} \\
&= Pr\{Y_1 = 1, Y_2 = 1\} \\
&= \pi^2.
\end{aligned}$$

Finally, the relationship between $Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}$ and $Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}$ can be explored as:

$$\begin{aligned}
&Pr\{Y_1 = 1, Y_2 = 0|R_2 > R_1\} \\
&= 1 - Pr\{Y_1 = 0, Y_2 = 0|R_2 > R_1\} - Pr\{Y_1 = 1, Y_2 = 1|R_2 > R_1\} - \\
&Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\} \\
&= 1 - (1 - \pi)^2 - \pi^2 - Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\}.
\end{aligned}$$

Then, we rewrite the Gini-like index as

$$\begin{aligned}
\text{Gini-like index} &= \frac{2Pr\{Y_1 = 0, Y_2 = 1|R_2 > R_1\} - 1 + (1 - \pi)^2 + \pi^2}{2E[Y]} \\
&= \frac{4\pi(1 - \pi)c - 1 + (1 - \pi)^2 + \pi^2}{2\pi} \\
&= (2c - 1)(1 - \pi).
\end{aligned}$$

Therefore, if $Y$ is a dichotomous outcome with $Pr\{Y = 1\} = \pi$, Gini-like index $= (2c - 1)(1 - \pi)$, where $2c - 1$ is commonly defined as the Kendall $\tau_k$ [40]. The Kendall's $\tau$ is a distribution-free measure for monotonic association for dependence between $Y$ and $R$.

## A.2 Equivalence of two $cfb$ definitions

The two definitions of $cfb$ are equivalent, which is shown as:

$$\begin{aligned}
cfb &= Pr\{(P_1 - P_2)(B_1 - B_2) > 0 | B_1 \neq B_2\} \\
&= Pr\{P_1 < P_2, B_1 < B_2 | B_1 \neq B_2\} + Pr\{P_2 > P_1, B_2 > B_1 | B_1 \neq B_2\} \\
&= 2Pr\{P_2 > P_1, B_2 > B_1 | B_1 \neq B_2\} \\
&= \frac{2Pr\{P_2 > P_1, B_2 > B_1, B_1 \neq B_2)}{Pr(B_1 \neq B_2)} \\
&= \frac{2Pr\{P_2 > P_1, B_2 > B_1)}{2Pr(B_1 > B_2)} \\
&= Pr\{P_2 > P_1 | B_2 > B_1).
\end{aligned}$$

Since the event of $B_1 > B_2$ and the event of $B_1 < B_2$ are mutually exclusive, $Pr\{B_1 \neq B_2\} = Pr\{B_1 > B_2\} + Pr\{B_1 < B_2\}$. Additionally, since $B_1$ and $B_2$ are independent and identical, $Pr\{B_1 > B_2\} = Pr\{B_1 > B_2\}$. Therefore, $Pr\{B_1 \neq B_2\} = 2Pr\{B_1 > B_2\}$.

## A.3 More about the $C_b$

To prove Equation 3.4, we first look at the numerator of the ratio in Equation 3.3. By the law of total expectation, $E[B] = E_X[E[B|X]]$. Then, we have

$$E[B] = E[P].$$

On the other hand, the denominator can be re-cast as a function of $P$ in the form of

$$E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)] = E[\max\{P_1, P_2\}]. \tag{A.1}$$

To prove Equation A.1, consider $B$ and $P$ as two functions of $X$. Also, $B_1 I(P_1 > P_2)$ and $B_2 I(P_1 < P_2)$ are expected values given a specific $X$. All together, we

have

$$
\begin{aligned}
&E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)] \\
&= E\big[E[B_1 I(P_1 > P_2)|X] + E[B_2 I(P_1 < P_2)|X]\big] \\
&= E\big[E[B_1|X] I(P_1 > P_2) + E[B_2|X] I(P_1 < P_2)\big] \\
&= E\big[P_1 I(P_1 > P_2) + P_2 I(P_1 < P_2)\big] \\
&= E\big[\max(P_1, P_2)\big].
\end{aligned}
$$

## A.4 More about estimator of $C_{b,h}$

This refers to Equation 3.14. Suppose we have a finite population with a size of $N$. Let $p_{h,i}$ be the predicted treatment effect of the $i$-th individual, where i = 1,2,3,..., N. Subjects in the population are ordered by $p_{h,r}$ in descending order, and $r$ represents the rank. Consequently, we have $p_{h,1} \geq p_{h,2} \geq ... \geq p_{h,N-1} \geq p_{h,N}$.

Then, $E[\max(P_{h,1}, P_{h,2})]$ can be written as the fraction, where the numerator is a weighted sum of $p_{h,r}$, and the denominator is the total number of pairs. Specifically, the weight of $p_{h,r}$ is the time to win the pairwise comparison among all possible pairs, and it is denoted as $t_r$. Then, the estimator can be expressed as the sum of all possible $t_r p_{h,r}$ over the total number of pairs, $n^2$. That is

$$
E[\max(P_{h,1}, P_{h,2})] = \frac{\sum_{r=1}^{N} t_r p_{h,r}}{N^2}. \tag{A.2}
$$

Based on the targeted treatment assignment rule, patients with highest ranking are more likely to win the pairwise comparison. The connection between $t_r$ and rank $r$ is $t_r = 2N - 2r + 1$. Then, we substitute $t_r$ in Equation A.2 and expand the

equation to get

$$
\begin{aligned}
E[\max(P_{h,1}, P_{h,2})] &= \frac{1}{N^2}\left\{\sum_{r=1}^{N}(2N-2r+1)p_{h,r}\right\} \\
&= \frac{1}{N^2}\left\{2\sum_{r=1}^{n}(N-r+1)p_{h,r} - \sum_{r=1}^{N}p_{h,r}\right\} \\
&= \frac{1}{n^2}\left\{2\sum_{r=1}^{N}\hat{S}(r) - \sum_{r=1}^{N}p_{h,r}\right\},
\end{aligned}
$$

where $\hat{S}(r)$ is the cumulative predicted benefit calculated as the partial sum of the predicted benefit up to and including the $r$-th element. We used $E[P_h] = \frac{1}{N}\sum_{r=1}^{N} p_{h,r}$ to estimate the expectation of the predicted treatment benefit $E[P_h]$. Therefore we have

$$
E[\max(P_{h,1}, P_{h,2})] = \frac{1}{N^2}\left\{2\sum_{r=1}^{N}\hat{S}(r) - \frac{1}{N}E[P_h]\right\}.
$$

The equation is shown in the parametric (model-based) estimator section of [12].

## A.5 The connection between $C_b$ and the Gini-like index for benefit

According to Equation 3.3 and Equation 3.6, we have

$$
\begin{aligned}
C_b &= 1 - \frac{E[B]}{E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)]}, \\
Gini_b &= \frac{E[B_1 - B_2 | P_1 > P_2]}{2E[B]}.
\end{aligned}
$$

$E[B_1 - B_2|P_1 > P_2]$ can be expressed as

$$\begin{aligned}
E[B_1 - B_2|P_1 > P_2] &= \frac{E[(B_1 - B_2)I(P_1 > P_2)]}{Pr(P_1 > P_2)} \\
&= 2E\big\{(B_1 - B_2)I(P_1 > P_2)\big\} \\
&= 2E\big\{B_1 I(P_1 > P_2) - B_2 I(P_1 > P_2)\big\} \\
&= 2E\big\{(B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2) \\
&- B_2 I(P_1 < P_2) - B_2 I(P_1 > P_2)\big\} \\
&= 2E\big\{(B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2) - E[B]\big\},
\end{aligned}$$

where $Pr(P_1 > P_2) = \frac{1}{2}$ is because $P$ is assumed to be continuous variable. Then Equation 3.6 can be rewritten as

$$\begin{aligned}
Gini_b &= \frac{E[B_1 - B_2|P_1 > P_2]}{2E[B]} \\
&= \frac{2E\big\{(B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2) - E[B]\big\}}{2E[B]} \\
&= \frac{2E\big\{(B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)\big\}}{2E[B]} - 1.
\end{aligned}$$

Therefore, we have

$$C_b = \frac{1}{Gini_b + 1}.$$

## A.6 Bivariate normal distributions in the stylized example

### A.6.1 The version with independent counterfactual outcomes

Recall that in the simple version we have

$$\begin{aligned}
X &\sim N(0, 1) \\
\varepsilon &\sim N(0, \sigma^2).
\end{aligned}$$

Thus, we have $P \sim N(\beta_0, \beta_1^2)$ and $B \sim N(\beta_0, \beta_1^2+\sigma^2)$. We assume that covariate $X$ and error $\varepsilon$ are independent. Thus, we have

$$\begin{aligned} Cov(X, \varepsilon) &= Cov(P, \varepsilon) \\ &= 0. \end{aligned}$$

As $P_1, P_2$ and $B_1, B_2$ are *iid* copies from the two-subject experiment, $P_2 - P_1 \sim N(0, 2\beta_1^2)$ and $B_2 - B_1 \sim N(0, 2(\beta_1^2+\sigma^2))$. The covariance of $P_2 - P_1, B_2 - B_1$ can be written as

$$\begin{aligned} cov(P_2 - P_1, B_2 - B_1) &= cov(P_2, B_2) + cov(P_1, B_1) \\ &= 2cov(P, B) \\ &= 2\left(cov(P, P) + cov(P, \varepsilon)\right) \\ &= 2\beta_1^2. \end{aligned}$$

Overall, we have

$$\begin{pmatrix} P_2 - P_1 \\ B_2 - B_1 \end{pmatrix} \sim N\Big( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2\beta_1^2 & 2\beta_1^2 \\ 2\beta_1^2 & 2\beta_1^2 + 2\sigma^2 \end{pmatrix} \Big).$$

### A.6.2 The version with dependent counterfactual outcomes

Continuing with the notation used in Section A.6.1, we compute $cfb$ as two times the cumulative distribution function of $(B_i - B_j)$ and $(P_i - P_j)$, which is denoted as $F(B_i - B_j, P_i - P_j; \mu, \Sigma)$. Now, as $(P_2 - P_1, B_2 - B_1)^T$ follows a centred bivariate normal distribution, we need to figure out the covariance matrix for $(P_2 - P_1, B_2 - B_1)^T$. Compared with the result in the previous version, the only difference is that the value of $Var(B_i - B_j)$ changes from $2\beta_1^2 + 2\sigma^2$ to

$$\begin{aligned} Var(B_i - B_j) &= 2Var(B) \\ &= 2\beta_1^2 + 4(1-\rho)\sigma^2. \end{aligned}$$

Therefore, joint distribution of $(P_2 - P_1, B_2 - B_1)^T$ in the counterfactual outcomes version is expressed as:

$$\begin{pmatrix} P_2 - P_1 \\ B_2 - B_1 \end{pmatrix} \sim N\Big( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2\beta_1^2 & 2\beta_1^2 \\ 2\beta_1^2 & 2\beta_1^2 + 4(1-\rho)\sigma^2 \end{pmatrix} \Big).$$

## A.7 More about $C_{b,h}$ in the linear set-up

For $C_b$, we have

$$\begin{aligned}
&1 - \frac{E[B]}{E[B_1 I(P_1 > P_2) + B_2 I(P_1 < P_2)]} = 1 - \frac{E[P]}{E[\max(P_1, P_2)]} \\
&= 1 - \frac{\beta_0}{\beta_0 + \beta_1 E[\max\{X_1, X_2\}]} \\
&= 1 - \frac{\beta_0}{\beta_0 + \frac{\beta_1}{\sqrt{\pi}}} \\
&= \frac{1}{\sqrt{\pi}\frac{\beta_0}{\beta_1} + 1},
\end{aligned}$$

In [41], an approach that computes the expectation of $\max\{X_1, X_2\}$ is provided. In general, we have $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. The first moment of $\max\{X_1, X_2\}$ can be calculated by differentiating the moment generating function of $\max\{X_1, X_2\}$. This provides

$$\begin{aligned}
E[\max\{X_1, X_2\}] &= \mu_1 \Phi\left(\frac{\mu_1 - \mu_2}{\theta}\right) + \mu_2 \Phi\left(\frac{\mu_2 - \mu_1}{\theta}\right) + \theta\phi\left(\frac{\mu_1 - \mu_2}{\theta}\right), \\
E[\min\{X_1, X_2\}] &= \mu_1 \Phi\left(\frac{\mu_2 - \mu_1}{\theta}\right) + \mu_2 \Phi\left(\frac{\mu_1 - \mu_2}{\theta}\right) - \theta\phi\left(\frac{\mu_2 - \mu_1}{\theta}\right),
\end{aligned}$$

where $\Phi(.)$ is the cumulative distribution function of $X$, $\phi(.)$ is the probability distribution function of $X$, $\theta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$, and $\rho$ is the correlation between $X_1$ and $X_2$.

Thus, when we have

$$\begin{aligned}
X_1 &\sim N(0, 1), \\
X_2 &\sim N(0, 1), \\
\rho &= 0,
\end{aligned}$$

the expectation of $\max\{X_1, X_2\}$ equals to

$$\begin{aligned} E[\max\{X_1, X_2\}] &= \sqrt{2}\frac{1}{\sqrt{2\pi}}e^0 \\ &= \frac{1}{\sqrt{\pi}}. \end{aligned}$$

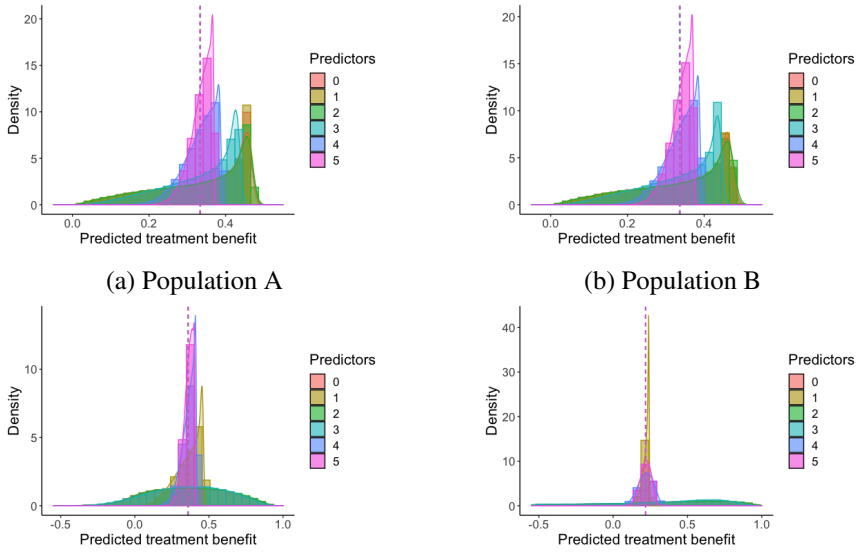
The expectation of $\min\{X_1, X_2\}$ equals to

$$\begin{aligned} E[\max\{X_1, X_2\}] &= -\sqrt{2}\frac{1}{\sqrt{2\pi}}e^0 \\ &= -\frac{1}{\sqrt{\pi}}. \end{aligned}$$

# Appendix B

# Figures and tables

This chapter consists of extra figures and tables of results.



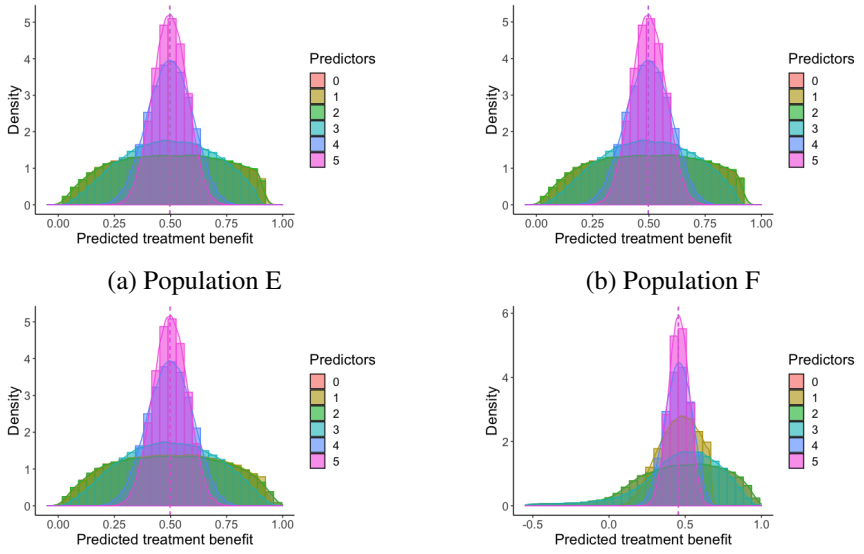

(a) Population A

(b) Population B

(c) Population C

(d) Population D

Figure B.1: Histograms of $P$ and $P_h$ for population A-D in Section 4.2, where the number zero represents $P$ and the number one to five represent five $P_h$, where $h = 1, 2, 3, 4, 5$, respectively.




(a) Population E

(b) Population F

(c) Population G

(d) Population H

Figure B.2: Histograms of $P$ and $P_h$ for population E-H in Section 4.2, where the number zero represents $P$ and the number one to five represent five $P_h$, where $h = 1, 2, 3, 4, 5$, respectively.

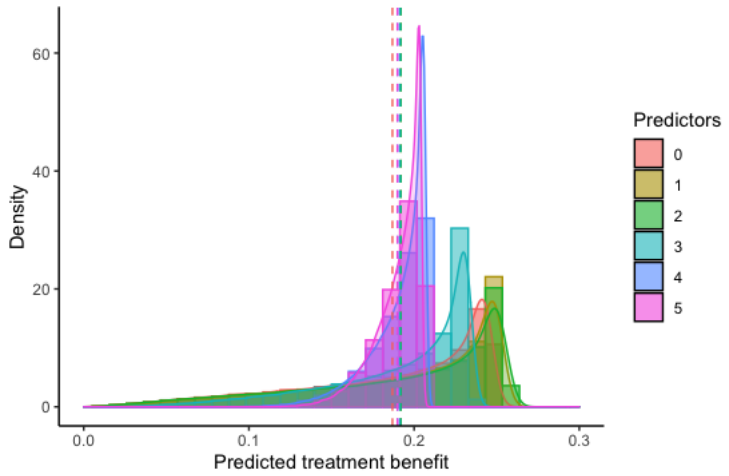



Figure B.3: Histogram of $P$ and $P_h$ for different levels of correlation among counterfactual outcomes in Section 4.3, where the number zero represents $P$ and the number one to five represent five $P_h$, where $h = 1, 2, 3, 4, 5$, respectively.

Table B.1: Population A-D and sets of model coefficients for benefit predictors

| *Population* | *Intercept* | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_a$ | $\hat{\beta}_{1,\Delta}$ | $\hat{\beta}_{2,\Delta}$ | $\hat{\beta}_{3,\Delta}$ |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| 1 | 0.006 | 1.004 | 0.504 | -0.405 | -2.014 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -2.016 | -0.001 | -0.009 | -0.028 |
| 3 | 0.007 | 0.923 | | | -1.875 | 0.023 | | |
| 4 | 0.003 | | 0.413 | | -1.664 | | 0.019 | |
| 5 | 0.009 | | | -0.312 | -1.654 | | | -0.051 |
| B | | | | | | | | |
| 1 | 0.006 | 0.987 | 0.506 | -0.405 | -2.030 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -2.021 | -0.044 | -0.004 | -0.027 |
| 3 | 0.007 | 0.923 | | | -1.878 | -0.017 | | |
| 4 | 0.003 | | 0.413 | | -1.688 | | 0.028 | |
| 5 | 0.009 | | | -0.312 | -1.676 | | | -0.054 |
| C | | | | | | | | |
| 1 | 0.005 | 0.609 | 0.486 | -0.374 | -2.134 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -2.031 | -1.006 | 0.014 | -0.003 |
| 3 | 0.007 | 0.923 | | | -1.880 | -0.923 | | |
| 4 | 0.003 | | 0.413 | | -1.969 | | 0.099 | |
| 5 | 0.009 | | | -0.312 | -1.937 | | | -0.075 |
| D | | | | | | | | |
| 1 | 0.006 | -0.212 | 0.342 | -0.264 | -0.984 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -2.027 | -4.015 | -0.008 | -0.010 |
| 3 | 0.007 | 0.923 | | | -1.902 | -3.745 | | |
| 4 | 0.003 | | 0.413 | | -1.688 | | 0.028 | |
| 5 | 0.009 | | | -0.312 | -0.946 | | | 0.121 |

Table B.2: Population E-H and sets of model coefficients for benefit predictors

| *Population* | *Intercept* | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_a$ | $\hat{\beta}_{1,\Delta}$ | $\hat{\beta}_{2,\Delta}$ | $\hat{\beta}_{3,\Delta}$ |
|---|---|---|---|---|---|---|---|---|
| E | | | | | | | | |
| 1 | 0.006 | 1.006 | 0.509 | -0.395 | -6.07 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -6.148 | 0.065 | 0.045 | 0.004 |
| 3 | 0.007 | 0.923 | | | -5.919 | 0.143 | | |
| 4 | 0.003 | | 0.413 | | -5.512 | | 0.138 | |
| 5 | 0.009 | | | -0.312 | -5.446 | | | -0.079 |
| F | | | | | | | | |
| 1 | 0.006 | 1.005 | 0.508 | -0.395 | -6.113 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -6.141 | 0.019 | 0.029 | 0.009 |
| 3 | 0.007 | 0.923 | | | -5.923 | 0.098 | | |
| 4 | 0.003 | | 0.413 | | -5.553 | | 0.123 | |
| 5 | 0.009 | | | -0.312 | -5.446 | | | -0.079 |
| G | | | | | | | | |
| 1 | 0.005 | 0.973 | 0.504 | -0.391 | -6.470 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -5.963 | -0.999 | -0.010 | 0.087 |
| 3 | 0.007 | 0.923 | | | -5.793 | -0.917 | | |
| 4 | 0.003 | | 0.413 | | -5.911 | | 0.084 | |
| 5 | 0.009 | | | -0.312 | -5.844 | | | 0.002 |
| H | | | | | | | | |
| 1 | 0.005 | 0.421 | 0.418 | -0.323 | -3.230 | | | |
| 2 | 0.006 | 1.004 | 0.508 | -0.395 | -6.002 | -4.013 | -0.014 | -0.012 |
| 3 | 0.007 | 0.923 | | | -5.729 | -3.816 | | |
| 4 | 0.003 | | 0.413 | | -3.068 | | -0.094 | |
| 5 | 0.009 | | | -0.312 | -3.058 | | | 0.060 |

Table B.3: The values of $(C_b, cfb)$, $(C_{b,h}, cfb_h)$ and $(\widehat{C}_{b,h}, \widehat{cfb_h})$ with different levels of correlation among covariates

| | | $\widehat{C}_{b,h}$ | | | | | $\widehat{cfb_h}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $h$ | $C_{b,h}$ | $Median$ | $Mean$ | $SD$ | $RMSE$ | $cfb_h$ | $Median$ | $Mean$ | $SD$ | $RMSE$ |
| $corr_x = 0$ | $C_b = 0.140$ and $cfb = 0.556$ | | | | | | | | | |
| 1 | 0.139 | 0.141 | 0.141 | 0.110 | 0.110 | 0.556 | 0.544 | 0.546 | 0.032 | 0.033 |
| 2 | 0.139 | 0.141 | 0.129 | 0.106 | 0.107 | 0.556 | 0.541 | 0.541 | 0.030 | 0.033 |
| 3 | 0.106 | 0.103 | 0.093 | 0.112 | 0.112 | 0.541 | 0.534 | 0.531 | 0.028 | 0.030 |
| 4 | 0.045 | 0.060 | 0.032 | 0.131 | 0.179 | 0.515 | 0.516 | 0.515 | 0.032 | 0.032 |
| 5 | 0.035 | 0.037 | 0.021 | 0.147 | 0.147 | 0.512 | 0.514 | 0.513 | 0.034 | 0.034 |
| $corr_x = 0.3$ | $C_b = 0.183$ and $cfb = 0.565$ | | | | | | | | | |
| 1 | 0.160 | 0.173 | 0.163 | 0.108 | 0.108 | 0.565 | 0.548 | 0.549 | 0.030 | 0.035 |
| 2 | 0.159 | 0.172 | 0.161 | 0.094 | 0.094 | 0.565 | 0.555 | 0.552 | 0.029 | 0.032 |
| 3 | 0.123 | 0.119 | 0.116 | 0.116 | 0.116 | 0.548 | 0.538 | 0.539 | 0.031 | 0.033 |
| 4 | 0.069 | 0.087 | 0.065 | 0.135 | 0.135 | 0.525 | 0.528 | 0.524 | 0.031 | 0.031 |
| 5 | 0.033 | 0.038 | 0.006 | 0.171 | 0.172 | 0.511 | 0.511 | 0.509 | 0.036 | 0.036 |
| $corr_x = 0.8$ | $C_b = 0.160$ and $cfb = 0.575$ | | | | | | | | | |
| 1 | 0.183 | 0.189 | 0.186 | 0.112 | 0.112 | 0.575 | 0.553 | 0.552 | 0.030 | 0.038 |
| 2 | 0.182 | 0.182 | 0.183 | 0.103 | 0.103 | 0.574 | 0.555 | 0.556 | 0.030 | 0.035 |
| 3 | 0.158 | 0.155 | 0.152 | 0.131 | 0.131 | 0.563 | 0.549 | 0.548 | 0.034 | 0.037 |
| 4 | 0.132 | 0.143 | 0.134 | 0.108 | 0.108 | 0.551 | 0.539 | 0.541 | 0.029 | 0.031 |
| 5 | 0.024 | 0.034 | -0.005 | 0.184 | 0.186 | 0.508 | 0.510 | 0.507 | 0.034 | 0.034 |