Statistica Sinica Preprint No: SS-2015-0174R3	
Title	Assessing the heterogeneity of treatment effects by identifying the treatment benefit and treatment harm ratio
Manuscript ID	SS-2015-0174R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0174
Complete List of Authors	Yunjian Yin Xiao-Hua Zhou Zhi Geng and Fang Lu
Corresponding Author	Xiao-Hua Zhou
E-mail	azhou@uw.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

ASSESSING THE HETEROGENEITY OF TREATMENT EFFECTS BY IDENTIFYING THE TREATMENT BENEFIT RATE AND TREATMENT HARM RATE

Yunjian Yin[1], Xiao-Hua Zhou[1,2,*], Zhi Geng[1,**], and Fang Lu[3]

[1]*School of Mathematical Sciences, Peking University, Beijing, China*

[2]*Department of Biostatistics, University of Washington, Seattle, USA*

[3]*China Academy of Chinese Medical Sciences*

Yunjian Yin: yinyunjian@pku.edu.cn; Xiao-Hua Zhou: azhou@uw.edu;

Zhi Geng: zhigeng@pku.edu.cn; Fang Lu: deerfang@126.com

[*]*Corresponding author phone:1-206-277-3588*

[**]*Corresponding author phone: 86-10-6245-1837*

Abstract: In a clinical trial, statistical reports have been typically concerned about the mean difference between two groups. Now there is increasing interest in the heterogeneity of the treatment effects, which means that the same treatment can have different effects on different people. In this article, we focus on the treatment benefit rate (TBR) and the treatment harm rate (THR), defined as the proportion of people who have a better outcome on the treatment than the control and the proportion of people who have a worse outcome on the treatment than the control,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

respectively. We propose a relatively weak assumption to obtain bounds for the TBR and the THR, which are shown to be always better than the covariates adjusted simple bounds. We prove that the TBR and THR are identifiable under a different conditional independence assumption. We also derive the corresponding estimators, the asymptotic distributions, and the over-identified test. We perform simulation studies to assess the performance of the proposed estimators and compare them with the proposed bounds. The simulation results show that the proposed estimators work quite well when the conditional independence assumption hold, they are not sensitive to small violation of the assumption, and the bounds we proposed can perform better than the estimators when the sample size is small. We illustrate application of the proposed methods in a double-blinded, randomized clinical trial.

Key words and phrases: Causal effect; Heterogeneity; Potential outcome; Treatment benefit rate; Treatment harm rate.

1 Introduction

In a typical phase III clinical trial, individuals are randomly assigned to either treatment or control, and then the relevant health endpoints are recorded; the difference between the means of the two study groups is used to estimate the average treatment effect (ATE). However, in general there are patients who do not benefit from an intervention even if the ATE is positive, and there are patients who can benefit from an intervention even if the ATE is significantly negative. Thus the ATE fails to capture variation in response to a treatment due to heterogeneity at many levels among patients in the target population (Davidoff (2009)).

It is important to understand the heterogeneity of treatment effects (HTE) in treatment

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

evaluation and selection. From a clinical perspective, many patients and healthcare providers may like to know not only the average effect, but also the proportion of people who have a worse outcome under the treatment than the control. In some situations, a treatment having a superior average effect may also have a greater risk of producing a deleterious effect for some patients. Understanding the HTE may also be important in forensic research (Gadbury, Iyer, and Allison (2001)). Subgroup analysis has been a common tool for understanding the HTE in the design of a clinical trial (Gail and Simon (1985), Russek-Cohen and Simon (1997), Pocock, Assmann, Enos, and Kasten (2002), Wang, Lagakos, Ware, Hunter, and Drazen (2007)), but it is perhaps more natural to study the HTE in terms of individual potential outcomes (Gadbury and Iyer (2000), Gadbury, Iyer, and Allison (2001), Gadbury, Iyer, and Albert (2004), Poulson, Gadbury, and Allison (2012)).

Some researchers have made extra assumptions regarding these proportions to resolve this heterogeneity. One of the main assumptions is "Monotonicity" (Goetghebeur and Molenberghs (1996) and Angrist, Imbens, and Rubin (1996)), which assumes that the treatment effect cannot be worse than the control for every individual. There are many scientific and empirical reasons to doubt this assumption. Berger, Rezvani, and Makarewicz (2003) suggested several explanations for the existence of individuals who would respond even to an inactive control but not to the experimental treatment, and pointed out that a placebo had been shown to be superior to an active treatment for some people.

Our main objective is to develop methods for studying the TBR and the THR based on the framework of potential outcomes (Rubin (1974), Rosenbaum and Rubin (1983), Holland (1986)). In this framework each patient is considered to have a potential outcome under each possible treatment, and the effect of an experimental treatment relative to a control can be assessed on each individual patient by comparing the corresponding potential outcomes. We focus on the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

case where the outcomes are binary variables. Because the TBR and the THR involve the joint distribution of the two potential outcomes from specific individuals that cannot be observed at the same time, they are not identifiable without additional assumptions, even in randomized clinical trials. Since simple bounds for them can be derived without further assumptions, various methods have been made to improve them, see Gadbury, Iyer, and Albert (2004); Albert, Gadbury, and Mascha (2005); Zhang, Wang, Nie, and Soon (2013).

Shen, Jeong, Li, Chen, and Buxton (2013) and Zhang, Wang, Nie, and Soon (2013) tried to identify THR by making the additional assumption that the two potential outcomes were independent conditional on observed covariates. In practice, this assumption is rarely true because it requires the two potential outcomes, which can not be observed at the same time and are always highly correlated, to be independent. Furthermore, it cannot be tested with the observed data. In addition, their method requires a specification of models between the covariates and the potential outcomes in estimation, which may lead to large bias when the models are misspecified.

We propose new methods to study the problem. First, we use the covariates to obtain bounds for the TBR and the THR. The covariates can be either discrete, continuous, or multidimensional, and the bounds we derive are always better than the adjusted simple bounds derived by Zhang, Wang, Nie, and Soon (2013). Then we identify the TBR and the THR under an assumption that requires at least three observed covariates to be conditionally independent. Under this assumption, we propose nonparametric estimators for the TBR and THR and derive the asymptotic distributions of the estimators. Compared to the estimators of Shen, Jeong, Li, Chen, and Buxton (2013) and Zhang, Wang, Nie, and Soon (2013), our estimators have two merits: the assumption for validation of our method can be tested with the observed data, and our estimators are nonparametric while their method requires parametric models.

We organize the paper as follows. The notation and assumptions are in Section 2. In Section

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

3 we give our bounds for the TBR and the THR. We derive the nonparametric estimators in Section 4. Simulation results are reported in Section 5. We also apply our methods to data analysis in Section 6. The paper ends with a discussion in Section 7.

2 Notation and Assumptions

Let T denote the binary assignment treatment variable (1 for treatment and 0 for control), and let Y stand for the binary outcome; $Y = 1$ if the subject survives or is cured, and $Y = 0$ if the subject dies or is not cured. We assume that a large value of Y indicates a better response. Let X be a set of covariates, which can be univariate or multivariate. We need two assumptions that are fundamental and widely used in causal inference.

Assumption 1. (Stable unit treatment value assumption, (SUTVA)) There is no interference between units: potential outcomes of one individual do not depend on the treatment status of other individuals and there is only one version of treatment (Rubin (1980)).

Under this assumption, we can denote Y_t as the potential outcome of a subject if the subject is assigned to treatment t, and $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$. Under the principal stratification framework (Frangakis and Rubin (2002)), we let G denote the principal stratum of an individual, defined as

$$G = \begin{cases} a, & Y_0 = 1, Y_1 = 1, \\ b, & Y_0 = 0, Y_1 = 1, \\ h, & Y_0 = 1, Y_1 = 0, \\ n, & Y_0 = 0, Y_1 = 0, \end{cases}$$

where “a”, “b”, “h” and “n” represent “always recover”, “benefit”, “harm” and “never recover”, respectively. The “benefit” stratum represents people who benefit from the treatment, and the “harm” stratum stands for people who suffer from it.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

Sheng, Jeong, Li, Chen, and Buxton (2013) defined the treatment benefit ratio (TBR) and the treatment harm ratio (THR) as

$$\text{TBR} : P(G = b) = P(Y_0 = 0, Y_1 = 1); \ \text{THR} : P(G = h) = P(Y_0 = 1, Y_1 = 0).$$

Assumption 2. (Randomization) $(X, G) \perp T$, or, $(X, Y_0, Y_1) \perp T$.

Under this assumption, we can identify the marginal distributions of Y_0 and Y_1, and then ATE $= E[Y_1 - Y_0]$ can be identified. With a little calculation,

$$\begin{cases} P(G = b) + P(G = n) = P(Y_0 = 0), & P(G = h) + P(G = n) = P(Y_1 = 0), \\ P(G = a) + P(G = b) + P(G = h) + P(G = n) = 1. \end{cases} \tag{1}$$

There are three equations and four parameters, so if one of the proportions of the four strata is identified or estimated, the others can also be identified or estimated.

3 Bounds based on closely related covariates

The TBR and the THR cannot be identified even in randomized trials without further assumptions. We derive bounds for these two rates. Let $p_1 = P(Y_1 = 1)$ and $p_0 = P(Y_0 = 1)$, which can be easily identified in a randomized trial. It is easy to get the bounds for the TBR and the THR:

$$\max(0, p_1 - p_0) \le \text{TBR} \le \min(1 - p_0, p_1), \quad \max(0, p_0 - p_1) \le \text{THR} \le \min(1 - p_1, p_0). \tag{2}$$

These bounds, referred to as simple bounds, do not need any further assumptions and can be easily estimated from the observed data. The bounds indicate that the TBR cannot be smaller than the ATE, which is equivalent to $p_1 - p_0$, and cannot be larger than the marginal probabilities

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

$P(Y_0 = 0)$ and $P(Y_1 = 1)$. Similarly, the THR cannot be smaller than the negative ATE, and cannot be larger than the marginal probabilities $P(Y_0 = 1)$ and $P(Y_1 = 0)$.

Zhang, Wang, Nie, and Soon (2013) used the covariates X to sharpen the bounds. With $p_{1X} = P(Y_1 = 1|X), p_{0X} = P(Y_0 = 1|X)$. their bounds after being adjusted by X are

$$\begin{aligned} &E\big[\max(0, p_{1X} - p_{0X})\big] \leq \text{TBR} \leq E\big[\min(1 - p_{0X}, p_{1X})\big], \\ &E\big[\max(0, p_{0X} - p_{1X})\big] \leq \text{THR} \leq E\big[\min(1 - p_{1X}, p_{0X})\big]. \end{aligned} \tag{3}$$

They pointed out that the adjusted bounds cannot be worse than the simple bounds.

We propose an assumption that can be used to further tighten the bounds.

Assumption 3. (Local Exclusion) Let S_0 and S_1 be two known subsets of the domain of X so that subjects with $X \in S_0$ do not fall in the "always recover" stratum, and subjects with $X \in S_1$ do not fall in the "never recover" stratum.

We call this "Local Exclusion" because we exclude one of the four strata defined by G in the subpopulation $X \in S_0$ and $X \in S_1$, while the "Monotonicity" assumption (Goetghebeur and Molenberghs (1996), Angrist, Imbens, and Rubin (1996)) excludes the "harm" stratum in the whole population.

To interpret the assumption, think of X as a variable that represents the severity of a disease. Here $X \in S_0$ if the subject has a serious disease, $X \in S_1$ if the subject has a mild disease, and X is outside S_0 and S_1 if the severity of the subject's disease is between serious and mild. Assumption 3 means that at least one of the two treatments cannot save the patient with serious disease, and at least one of the two treatment can save the patient with mild disease. Take drug therapy aiming at helping patients recover from bacterial infection inflammation as an example. Take $T = 1$ if the individual receives the drug treatment, and $T = 0$ if the individual is assigned to placebo-treated. Let $Y = 1$ if the individual is cured, $Y = 0$ if not. Let X be the indicator variable, representing the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

severity of inflammation with large values meaning severe inflammation. People with very serious inflammation cannot recover from the placebo treatment, while people with mild inflammation can be cured by the drug treatment. Thus, the individual with large value of X cannot be in the "always recover" group ($G = a$) and the individual with small value of X can not be in the "never recover" group ($G = n$). Here $X \in S_0$ means X has a large value while $X \in S_1$ means the value of X is small.

With Assumption 3, one can have $P(G = a|X \in S_0) = P(G = n|X \in S_1) = 0$. From (1) we can conclude that the joint distribution of (Y_0, Y_1) conditional on $X \in S_k, k = 0, 1$ can be identified. Thus, the TBR and the THR conditional on $X \in S_k, k = 0, 1$ can also be identified, as

$$
\begin{aligned}
&P(G = b|X \in S_0) = P(G \in \{a, b\}|X \in S_0) = P(Y_1 = 1|X \in S_0) = P(Y = 1|X \in S_0, T = 1),\\
&P(G = b|X \in S_1) = P(G \in \{b, n\}|X \in S_1) = P(Y_0 = 0|X \in S_1) = P(Y = 1|X \in S_1, T = 0),\\
&P(G = h|X \in S_0) = P(G \in \{a, h\}|X \in S_0) = P(Y_0 = 1|X \in S_0) = P(Y = 1|X \in S_0, T = 0),\\
&P(G = h|X \in S_1) = P(G \in \{h, n\}|X \in S_1) = P(Y_1 = 0|X \in S_1) = P(Y = 1|X \in S_1, T = 1).
\end{aligned}
$$

Let $S_2 = \overline{S_0 \bigcup S_1}$, where $\overline{\cdot}$ stands for the complement operation. For the TBR and the THR in the subpopulation with $X \in S_2$, bounds can be obtained by adjusting simple bounds with X as in (3).

Theorem 1. Under Assumptions 1, 2, and 3, we have

$$
\begin{aligned}
\text{TBR} &\geq L_b = P(Y_1 = 1, X \in S_0) + P(Y_0 = 0, X \in S_1) + E[\max(0, p_{1X} - p_{0X})I(X \in S_2)],\\
\text{TBR} &\leq U_b = P(Y_1 = 1, X \in S_0) + P(Y_0 = 0, X \in S_1) + E[\min(p_{1X}, 1 - p_{0X})I(X \in S_2)],\\
\text{THR} &\geq L_h = P(Y_0 = 1, X \in S_0) + P(Y_1 = 0, X \in S_1) + E[\max(0, p_{0X} - p_{1X})I(X \in S_2)],\\
\text{THR} &\leq U_h = P(Y_0 = 1, X \in S_0) + P(Y_1 = 0, X \in S_1) + E[\min(p_{0X}, 1 - p_{1X})I(X \in S_2)].
\end{aligned}
$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

We denote these as "LE" bounds (Local Exclusion). Here (L_b, U_b, L_h, U_h) can be identified due to Assumption 2. The widths of the two bounds depend largely on $P(X \in S_2)$. The smaller the probability is, the narrower the widths are. Moreover, the TBR and the THR become identifiable when $P(X \in S_2) = 0$.

The "LE" bounds are better than the covariates adjusted bounds in (3), since we make full use of the information about the relationship between X and G in the "LE" bounds. The proof is in the supplementary materials.

Proposition 1. The "LE" bounds for the TBR and the THR are no worse than the bounds in (3); they are equivalent if and only if $P(X \in S_0) + P(X \in S_1) = 0$.

If X is a discrete variable, the bounds can be easily estimated by the moment estimator. Let $\widehat{L}_h, \widehat{U}_h, \widehat{L}_b, \widehat{U}_b$ be the resulting non-parametric estimators for the lower and upper bounds of the TBR and the THR, respectively. They have the form

$$\begin{aligned}
\widehat{L}_h &= \frac{P_n[f(1,0,0)]}{P_n[I(T=0)]} + \frac{P_n[f(0,1,1)]}{P_n[I(T=1)]} + \sum_{x\in S_2} \max\Big\{0, \frac{P_n[g(1,x,0)]}{P_n[I(T=0)]} - \frac{P_n[g(1,x,1)]}{P_n[I(T=1)]}\Big\}, \\
\widehat{U}_h &= \frac{P_n[f(1,0,0)]}{P_n[I(T=0)]} + \frac{P_n[f(0,1,1)]}{P_n[I(T=1)]} + \sum_{x\in S_2} \min\Big\{\frac{P_n[g(1,x,0)]}{P_n[I(T=0)]}, \frac{P_n[g(0,x,1)]}{P_n[I(T=1)]}\Big\}, \\
\widehat{L}_b &= \frac{P_n[f(1,0,1)]}{P_n[I(T=1)]} - \frac{P_n[f(0,1,0)]}{P_n[I(T=0)]} + \sum_{x\in S_2} \max\Big\{0, \frac{P_n[g(1,x,1)]}{P_n[I(T=1)]} - \frac{P_n[g(1,x,0)]}{P_n[I(T=0)]}\Big\}, \\
\widehat{U}_b &= \frac{P_n[f(1,0,1)]}{P_n[I(T=1)]} + \frac{P_n[f(0,1,0)]}{P_n[I(T=0)]} + \sum_{x\in S_2} \min\Big\{\frac{P_n[g(0,x,0)]}{P_n[I(T=0)]}, \frac{P_n[g(1,x,1)]}{P_n[I(T=1)]}\Big\},
\end{aligned}$$

where $P_n[\cdot]$ is the empirical mean, $I(\cdot)$ is the indicator function, $f(j_1, j_2, j_3) = I(Y = j_1, X \in S_{j_2}, T = j_3)$, and $g(\ell_1, x, \ell_2) = I(Y = \ell_1, X = x, T = \ell_2)$.

We use the percentile bootstrap method to construct confidence intervals for the lower bounds and upper bounds: we randomly draw datasets from the original sample with replacement, and with the new dataset, we compute estimates of (L_b, U_b, L_h, U_h), denoted as $(\widehat{L}_b^*, \widehat{U}_b^*, \widehat{L}_h^*, \widehat{U}_h^*)$. The process is repeated B times to get $\big((\widehat{L}_{b,1}^*, \widehat{U}_{b,1}^*, \widehat{L}_{h,1}^*, \widehat{U}_{h,1}^*), \cdots, (\widehat{L}_{b,B}^*, \widehat{U}_{b,B}^*, \widehat{L}_{h,B}^*, \widehat{U}_{h,B}^*)\big)$. We form

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

approximate 95% confidence intervals by finding the 2.5% and 97.5% percentiles of $(\widehat{L}_b^*, \widehat{U}_b^*, \widehat{L}_h^*, \widehat{U}_h^*)$, denoted as $(\widehat{L}_{b,(2.5)}^*, \widehat{U}_{b,(2.5)}^*, \widehat{L}_{h,(2.5)}^*, \widehat{U}_{h,(2.5)}^*)$ and $(\widehat{L}_{b,(97.5)}^*, \widehat{U}_{b,(97.5)}^*, \widehat{L}_{h,(97.5)}^*, \widehat{U}_{h,(97.5)}^*)$, respectively. Then the approximate 95% confidence intervals for the bounds of the TBR and the THR can be constructed as $[\widehat{L}_{b,(2.5)}^*, \widehat{U}_{b,(97.5)}^*]$ and $[\widehat{L}_{h,(2.5)}^*, \widehat{U}_{h,(97.5)}^*]$.

4 Nonparametric Identifiability and Estimation

In this Section, we first consider the nonparametric identification of the TBR and the THR under an assumption, then derive the nonparametric estimators, the asymptotic distributions, and the over-identified test.

4.1 Nonparametric Identifiability

Let X be a vector of observed covariates, where $X = (X_1, ..., X_k)$. We assume that $X_j, j = 1, .., k$, are binary variables. The assumption is for convenience and is not necessary.

Assumption 4. $X_1, ..., X_k$ are mutually independent in the "always recover" group and "never recover" group.

Assumption 4 can be true with properly chosen $X_1, \cdots, X_k$ in some settings. Consider treatment as a medicine or a therapy aimed at curing a certain kind of disease with a binary outcome: whether a patient is cured (1 if cured and 0 if not), and then we can choose the covariates X as some of the symptoms. The patients' symptoms are not mutually independent but can reflect a latent common cause (Elrington, Murray, Spiro, and Newsom-Davis (1991)). The disease, of course, is the cause. The symptoms are likely to be mutually independent given the common cause (disease). Thus we can assume that some symptoms are mutually independent in the serious disease class and the slight disease class. Here $G = a$ $(G = n)$ means that, regardless of what treatment the patient receives, he/she would be cured (still suffer from the disease) at the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

end of the study. If someone gets a serious disease, neither of the treatments can save him/her from the disease; alternatively, if someone gets a slight disease, he/she would be cured under either of the two treatments. So $G = a$ can represent slight disease and $G = n$ can represent serious disease. It is reasonable to assume that some symptoms are mutually independent in the strata $G = a$ and $G = n$.

In another example, we consider a hypothetical randomized clinical trial of a new drug against a placebo for treating a disease. Let $X_1, \cdots, X_k$ be the diagnosis on the severity of the disease by k different doctors with different medical backgrounds. Given the true but latent severity level of the disease, $X_1, \cdots, X_k$ could be conditionally independent because the k doctors make their diagnoses on the severity of disease based on their own experiences rather than any other common variables. Furthermore, since the group with $G = a$ (i.e. $Y_0 = Y_1 = 1$) consists of patients with lightly severe disease, and the group with $G = n$ (i.e. $Y_0 = Y_1 = 0$) consists of patients with severe disease. It is reasonable to believe that $X_1, \cdots, X_k$ are independent conditional on $G = a, n$. Therefore, Assumption 4 holds for the chosen variables, $X_1, \cdots, X_k$.

This kind of assumption has been used in other settings. The naive Bayes classifier uses the assumption that the covariates are mutually independent given the true class, and it can classify individuals quite well in many applications (Bickel and Levina (2004)). Similarly, it is generally assumed that the observed variables are mutually independent within clusters for dealing with unobserved heterogeneity in latent class analysis (Vermunt and Magidson (2002)). In estimation of the accuracy of diagnostic tests, it is usually assumed that the test results are independent conditional on the unobserved disease status (Zhou, McClish, and Obuchowski (2012, Chap.11)). In general, if we choose covariates that are manifestation of a latent variable that is highly related to the $G = n$ and $G = a$ groups, then Assumption 4 most likely holds.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

Under Assumption 4, we have

$$\begin{aligned} P(X_1,...,X_k,Y=1|T=1) &= P(X_1,...,X_k|G=a)\pi_a + P(X_1,...,X_k|G=b,T=1,Y=1)\pi_b \\ &= P(X_1|G=a)...P(X_k|G=a)\pi_a + P(X_1,...,X_k|G=b)\pi_b, \end{aligned}$$

where $\pi_g = p(G=g), g=a,b,h,n$. Similarly, we have

$$P(X_1,\cdots,X_k,Y=0|T=0) = P(X_1|G=n)\cdots P(X_k|G=n)\pi_n + P(X_1,\cdots,X_k|G=b)\pi_b.$$

We want to identify $\pi = (\pi_a,\pi_b,\pi_h,\pi_n)$. By rearranging these equations to eliminate some nuisance parameters, we have

$$\begin{aligned} &P(X_1,\cdots,X_k,Y=1|T=1) - P(X_1,\cdots,X_k,Y=0|T=0) \\ &= P(X_1|G=a)\cdots P(X_k|G=a)\pi_a - P(X_1|G=n)\cdots P(X_k|G=n)\pi_n. \end{aligned} \tag{4}$$

There are 2^k equations, and 2(k+1) parameters here, hence to identify the parameters we need $2^k \geq 2(k+1)$, i.e., $k \geq 3$. Having more equations than parameters is not enough to guarantee an unique solution, we need the following assumption:

Assumption 5. There exists at least one covariate in $\{X_1,...,X_k\}$, say X_j, such that $P(X_j|G=a) \neq P(X_j|G=n)$.

A proof of the following is given in the supplementary materials.

Theorem 2. When $k \geq 3$, if Assumptions 1, 2, 4 and 5 hold, the TBR and the THR are identifiable.

Thus if there are at least three covariates $\{X_1,\cdots,X_k\}$ that are independent conditional on $G=a$ and $G=n$, the TBR and the THR are identifiable when Assumption 5 holds true.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

4.2 Nonparametric estimation

Our nonparametric estimators are based on the generalized method of moments (GMM) estimator as formalized by Hansen (1982). For simplicity, we assume all $X_1, \cdot, X_k$ are binary variables. The non-binary case is discussed later.

Let $\rho_{aj} = P(X_j = 1|G = a), \rho_{nj} = P(X_j = 1|G = n), \pi_g = P(G = g), g = a, b, h, n, p_1 = P(T = 1), p_2 = P(Y = 1|T = 1)$, and $\theta = \{\pi_b, \pi_h, \rho_{a1}, \cdots, \rho_{ak}, \rho_{n1}, \cdots, \rho_{nk}, p_1, p_2\}$. By substituting $\pi_a = p_2 - \pi_b, \pi_n = 1 - p_2 - \pi_h$ into (4), we have

$$\begin{aligned} & P(X_1 = x_1, ..., X_k = x_k, Y = 1|T = 1) - P(X_1 = x_1, ..., X_k = x_k, Y = 0|T = 0) \\ = \quad & (p_2 - \pi_b)\varphi_{a1}(x_1) \cdots \varphi_{ak}(x_k) - (1 - p_2 - \pi_h)\varphi_{n1}(x_1) \cdots \varphi_{nk}(x_k), \end{aligned}$$

where $\varphi_{gj}(x_j) = \rho_{gj}^{x_j}(1 - \rho_{gj})^{1-x_j}, g = a, n, j = 1, \cdots, k$. Let

$$g_1(\theta) = P_n[\widetilde{g}_1(\theta)] = P_n\big[I(T = 1) - p_1\big],$$

$$g_2(\theta) = P_n[\widetilde{g}_2(\theta)] = P_n\big[I(Y = 1, T = 0) - (1 - p_1)p_2\big],$$

$$\begin{aligned} & g(x_1, \cdots, x_k; \theta) \\ = \quad & P_n[\widetilde{g}(x_1, \cdots, x_k; \theta)] \\ = \quad & P_n\Big[\Big(I(X_1 = x_1, ..., X_k = x_k, Y = 1, T = 1)/p_1 - I(X_1 = x_1, ..., X_k = x_k, Y = 0, T = 0)/(1 - p_1)\Big) \\ & - \Big((p_2 - \pi_h)\varphi_{a1}(x_1) \cdots \varphi_{ak}(x_k) - (1 - p_2 - \pi_b)\varphi_{n1}(x_1) \cdots \varphi_{nk}(x_k)\Big)\Big], \end{aligned}$$

$$g(\theta) = \Big(g_1(\theta), g_2(\theta), g(x_1, \cdots, x_k; \theta), x_j \in \{0, 1\}, j = 1, \cdots, k\Big)^T,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

where $I(\cdot)$ is the indicator function. Then the GMM estimator $\widehat{\theta}_n$ is

$$\widehat{\theta}_n = \arg\min_{\theta} Q(\theta) = \arg\min_{\theta} g(\theta)^T W(\theta)^{-1} g(\theta), \tag{5}$$

where $W(\theta)$ is a positive semi-definite matrix. To reduce the computational burden, we use a two-step procedure to estimate θ. Estimates are constructed by using a preliminary weighting matrix $\widehat{W}$ (the identify matrix is used here) to replace $W(\theta)$ in (5), and we take $\widehat{\theta}_{n1}$ to be a solution to the initial optimization problem,

$$G(\widehat{\theta}_{n1})^T \widehat{W}^{-1} g(\widehat{\theta}_{n1}) = 0,$$

where $G(\theta) = \frac{\partial}{\partial \theta} g(\theta)$. If $S(\theta)$ is the sample covariance matrix of $g(\theta)$, the estimator $\widehat{\theta}_n$ is defined as in (5) by replacing $W(\theta)$ with $S(\widehat{\theta}_{n1})$, specifically,

$$G(\widehat{\theta}_n)^T S(\widehat{\theta}_{n1})^{-1} g(\widehat{\theta}_n) = 0.$$

By the theory of GMM, the estimator has the following asymptotic property.

Theorem 3. Under Assumptions 1, 2, 4, 5 with $k \geq 3$,

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \xrightarrow{d} N\big(0, (G^T S^{-1} G)^{-1}\big),$$

where $\xrightarrow{d}$ means convergence in distribution, and $G = \lim_n G(\theta), S = \lim_n S(\theta)$. In particular, the estimator $(\widehat{\pi}_b, \widehat{\pi}_h)$ of (π_b, π_h) satisfies

$$\sqrt{n}\Big(\begin{pmatrix} \widehat{\pi}_b \\ \\ \widehat{\pi}_h \end{pmatrix} - \begin{pmatrix} \pi_b \\ \\ \pi_h \end{pmatrix} \Big) \xrightarrow{d} N\big(0, \Sigma\big),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

where Σ is the corresponding 2×2 block matrix of $(G^T S^{-1} G)^{-1}$.

For simplicity, we call $(\widehat{\pi}_b, \widehat{\pi}_h)$ the "CI" estimators (conditionally independent) and Assumption 4 with $k \geq 3$ the "CI" assumption.

The variance $(G^T S^{-1} G)^{-1}$ can be estimated by $(G^T(\widehat{\theta}_n) S(\widehat{\theta}_n)^{-1} G(\widehat{\theta}_n))^{-1}$. So we have an estimator for Σ, denoted by $\widehat{\Sigma}$. The 95% confidence intervals for π_b and π_h can then be constructed as

$$\big[\widehat{\pi}_b - 1.96\sqrt{\widehat{\Sigma}_{11}},\ \widehat{\pi}_b + 1.96\sqrt{\widehat{\Sigma}_{11}}\big],\ \big[\widehat{\pi}_h - 1.96\sqrt{\widehat{\Sigma}_{22}},\ \widehat{\pi}_h + 1.96\sqrt{\widehat{\Sigma}_{22}}\big],$$

where $\widehat{\Sigma}_{ij}$ is the corresponding element in the matrix $\widehat{\Sigma}$.

4.3 Over-identified test and backward variables selection

In practice, one can question the validity of Assumption 4. The GMM method provides a test, the over-identified or J-test, when we have more than three covariates. The J-statistics is

$$J = n g(\widehat{\theta}_n)^T W(\widehat{\theta}_n)^{-1} g(\widehat{\theta}_n) \rightsquigarrow \chi^2(2^k - 2k - 2).$$

When the p-value of the proposed J-statistics is smaller than a pre-specified significant level, usually 0.05, we can reject Assumption 4.

In general, many symptoms may be collected in a clinical trial. We can use the backward selecting method with the J-test to select appropriate ones. The covariate selection procedure is as follows:

---

1. Initialize $X_{\text{new}} = X$.

2. Calculate the J-test statistics and its p-value with the covariates X_{new}; if the p-value is bigger than 0.05 or $\dim(X_{\text{new}}) \leq 4$, then stop; if not, go to step 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

3. Remove the r-th component $X_{\text{new},r}$ from X_{new}, saying $X_{\text{new},-r}$, where $r = 1, \cdots, \dim(x_{\text{new}})$, and calculate the corresponding J-statistics J_r and p-value p_r; update

$$X_{\text{new}} = X_{\text{new},-\widetilde{r}}, \text{ where } \widetilde{r} = \arg\max_r p_r.$$

Then go back to step 2.

If the finial p-value is smaller than 0.05, then Assumption 4 may not hold.

5 Simulation Studies

In this section, we report the results of two simulation studies. We evaluated the performances of the "CI" estimators when the "CI" assumption holds and does not hold. The performance was measured by bias, and bias percentage, which was defined by $100 \times \big|\text{bias/true value}\big|\%$. We also estimated the average asymptotic standard error (ASE), the empirical standard error (ESE), and the coverage of 95% confidence intervals. We also compare the average length of the confidence intervals (ALCIs) of the "LE" bounds and the "CI" estimators under different sample sizes.

In the first simulation study, we generated 1000 samples for several independent variables: $(T, G, \xi, \xi_k, k = 1, 2, 3, 4)$. Here T was the binary treatment assignment with $P(T = 1) = 0.5$, and G was the principal stratum, which followed a multinomial distribution with the cell probability $\big\{P(G = g), g = a, b, h, n\big\} = (0.4, 0.3, 0.2, 0.1)$. According to the definition of G, both potential outcomes, Y_0 and Y_1, are determined once G is determined; and ξ, ξ_1, ξ_2, ξ_3, and ξ_4 were generated independently from the standard normal distribution. We then constructed the covariates as

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

follows:

$$\begin{cases} \text{In the subgroup } G = g,\ \widetilde{X}_k = \mu_g + \alpha_{g,k}\xi + \xi_k,\ k = 1, 2, 3, 4,\ g = a, b, h, n, \\ X_k = I(\widetilde{X}_k > 0). \end{cases}$$

We set $\mu = (\mu_a, \mu_b, \mu_h, \mu_n) = (1, 0.3, -0.4, -1), \alpha_b = (\alpha_{b,1}, \alpha_{b,2}, \alpha_{b,3}, \alpha_{b,4}) = (1.5, -1, 1, -1.2)$, $\alpha_h = (\alpha_{h,1}, \alpha_{h,2}, \alpha_{h,3}, \alpha_{h,4}) = (-1.2, 1, 0.5, -2)$, and $\alpha_a = (\alpha_{a,1}, \alpha_{a,2}, \alpha_{a,3}, \alpha_{a,4}) = \alpha_n = (\alpha_{n,1}, \alpha_{n,2}, \alpha_{n,3}, \alpha_{n,4}) = \gamma \cdot (1, 1, 1, 1)$. The "CI" assumption holds when $\gamma = 0$. As γ increases from 0, the correlation between $(X_1, \cdots, X_4)$ conditional on $G = a, n$ increases. Larger values of γ can cause large violations of Assumption 4, which can induce large bias in the "CI" estimators.

With each data set, we calculated the J-test statistics and the corresponding p-value. If the p-value was smaller than 0.05, the "CI" assumption was rejected. We only used the data sets for which the p-values were greater than 0.05 to assess the performance of the "CI" estimators. These results are reported in Table 1.

Based on the results in Table 1, we can draw the following conclusions. When the "CI" assumption holds, $\gamma = 0$, the estimators perform very well with small bias and bias-percentage. The rejection rate is almost 5%, which means the J-test also performs very well. The ASE is approximated to the ESE and the coverage is nearly 95%, which means the estimators of the variances also work quite well.

When the "CI" assumption is violated, the power of the J-test increases as γ increases. When $\gamma < 0.5$, the bias and the bias percentage are still small and the coverage of the 95% confidence intervals is almost 95%. Thus, our estimators are not sensitive to small violation of the assumption here. When γ continues to increase ($\gamma > 0.5$), the bias and the bias percentage do not increase a lot. The ASE is always approximately equal to the ESE, which means the estimators of the variances perform very well even when the assumption is violated. However, when $\gamma > 0.7$, the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

Table 1: The performance of our estimators, J-test and the coverage of 95% confidence intervals

γ	rejection rate	bias percentage	bias	ASE	ESE	coverage
0.0	0.053	0.020	-0.006	0.020	0.020	0.944
		0.011	-0.002	0.026	0.026	0.953
0.1	0.056	0.020	-0.006	0.020	0.019	0.950
		0.008	-0.002	0.026	0.026	0.948
0.2	0.061	0.016	-0.005	0.020	0.020	0.954
		0.006	-0.001	0.026	0.026	0.951
0.3	0.068	0.010	-0.003	0.020	0.020	0.953
		0.002	0.000	0.026	0.026	0.956
0.4	0.072	0.000	0.000	0.021	0.020	0.959
		0.014	0.003	0.026	0.026	0.950
0.5	0.074	0.012	0.004	0.021	0.020	0.956
		0.034	0.007	0.026	0.025	0.945
0.6	0.088	0.028	0.009	0.021	0.020	0.940
		0.056	0.011	0.026	0.025	0.934
0.7	0.097	0.048	0.014	0.021	0.020	0.914
		0.085	0.017	0.026	0.025	0.909
0.8	0.127	0.069	0.021	0.021	0.020	0.863
		0.115	0.023	0.026	0.025	0.874
0.9	0.155	0.090	0.027	0.021	0.020	0.781
		0.144	0.029	0.026	0.024	0.823
1.0	0.215	0.113	0.034	0.022	0.019	0.676
		0.174	0.035	0.026	0.024	0.755

The two elements in some table cells correspond to the TBR (first row in each cell) and the THR (second row in each cell), respectively.

coverage of the 95% confidence intervals decreases rapidly.

In the second simulation study, we compared the performances of the "LE" bounds and the "CI" estimators under different sample sizes. The simulation study was conducted in following steps:

Step 1: A set of values for the sample size n was created. Variables T and G were generated independently: T was generated from a Bernoulli distribution with $P(T = 1) = 0.5$, G was generated by randomly drawing from $\{a, b, h, n\}$ with probabilities $(0.4, 0, 3, 0.2, 0.1)$. With T and G generated, the outcome Y was decided by the definition of G. The covariate X_{LE}

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

that was used in obtaining the "LE" bounds and the covariates $X_{\rm CI} = (X_{\rm CI,1}, \cdots, X_{\rm CI,4})$ that was used in "CI" estimators were generated independently given G. The distribution of $X_{\rm LE}$ was as follows:

$$P(X_{\rm LE} = 2|G = a) = 1/2, P(X_{\rm LE} = 3|G = a) = 1/2,$$

$$P(X_{\rm LE} = 1|G = n) = 1/2, P(X_{\rm LE} = 2|G = n) = 1/2,$$

$$P(X_{\rm LE} = 1|G = g) = 1/3, P(X_{\rm LE} = 2|G = g) = 1/3, P(X_{LE} = 3|G = g) = 1/3, g = b, h.$$

Assumption 3 is valid by setting $S_0 = \{1\}, S_1 = \{3\}$. The four components of $X_{\rm CI}$ were all binary and were generated independently in the subgroups $G = a, n$ with the probabilities:

$$P(X_{{\rm CI},k} = 1|G = a) = 0.8, P(X_{{\rm CI},k} = 1|G = n) = 0.2, k = 1, 2, 3, 4.$$

In the subgroups $G = b, h$, we first generated $(\epsilon, \epsilon_1, \cdots, \epsilon_4)$, mutually independent and standard normal. The covariates $X_{\rm CI}$ were constructed as,

$$\text{In subgroup } G = g, X_{{\rm CI},k} = I(\mu_{g,k} + \epsilon + \epsilon_k > 0), g = b, h, k = 1, 2, 3, 4,$$

where $\mu_b = (\mu_{b,1}, \cdots, \mu_{b,4}) = (-1, -0.4, 0.3, 1)$, $\mu_h = (\mu_{h,1}, \cdots, \mu_{h,4}) = (1, 0.3, -0.4, -1)$.

Step 2: With the data $\{T, Y, X_{\rm LE}\}$, we estimated the "LE" bounds for the TBR and the THR. The 95% confidence intervals were estimated by the bootstrap method described in Section 3. With the data $\{T, Y, X_{\rm CI}\}$, we obtained the "CI" estimates and the confidence intervals for the TBR and the THR.

Step 3: Step 1 and Step 2 were repeated 1000 times to estimate the ALCIs of "LE" bounds and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

"CI" estimators.

Figure 1 shows the ALCIs for the TBR and the THR under different sample sizes. We can see that, when the sample size is small, the ALCIs of the "CI" estimators can be wider than the "LE" bounds. As the sample size increases, the ALCIs of both methods decrease, but the "CI" estimators method is faster. When the sample size passes a certain threshold, the ALCIs of the "CI" estimators are smaller than the "LE" bounds.

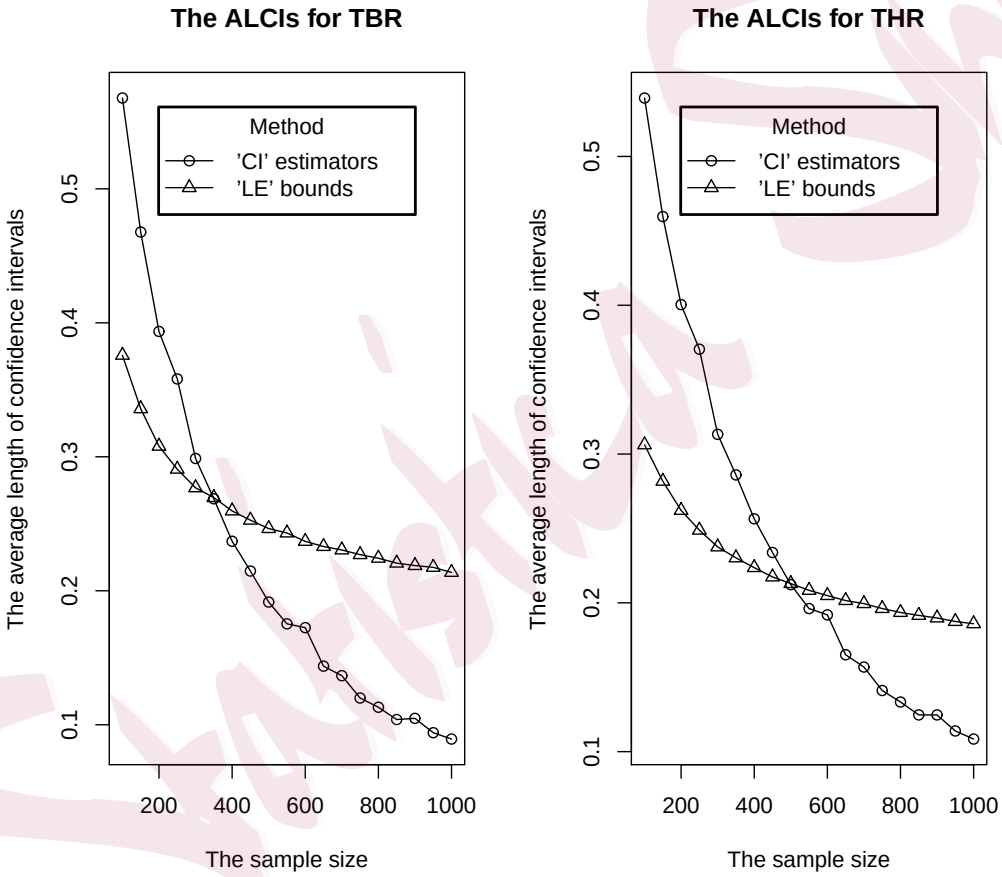

Figure 1: The comparison of the length of the confidence intervals for CI estimators and LE bounds

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

6 Analysis of a randomized clinical trial

We applied the proposed methods to estimate the TBR and the THR of a drug that treats acute bronchitis. The data was from a randomized, double-blind, placebo-controlled clinical trial. In the original study, subjects were assigned to one of three groups: the high dose group, the low dose group and the placebo group. To illustrate the proposed method, we focused on the effect of the high dose ("treatment") versus placebo. The study sample consisted of 155 patients with acute bronchitis. The outcome of interest was the sum of the scores for three ordinal-scale symptoms: the cough (0 for no cough, 1 for a small cough, 2 for frequent cough which mildly affects the daily activities, 3 for frequent cough which seriously affects the daily activities), the amount of expectoration (0 for less than 10ml a day, 1 for between 10ml and 50ml a day, 2 for between 50ml and 100ml a day, 3 for more than 100ml a day), and the quality of expectoration (0 for none, 1 for white expectoration and easily coughed up, 1 for yellowish and hard to cough up, 2 for yellow and hard to cough up). Let Z_1 and Z_2 denote the sum of three symptom scores at baseline and the end of the trial, respectively, and $Z = (Z_1 - Z_2)/Z_1$, representing the percentage decline relative to the baseline; since Z_1 is always strictly bigger than 0 in the trial, Z is well defined. If $Z > 70\%$, the drug is considered effective. We focused on $Y = I(Z > 70\%)$, where $I(\cdot)$ is the indicator function. We considered the individual to be cured if $Y = 1$, and not cured if $Y = 0$. The randomization in this trial can be used to estimate the ATE, which is estimated at 0.472. The treatment has a better average effect than the placebo. But there may still exist individuals who are harmed by the treatment. We applied the methods proposed in this paper to obtain the bounds and estimations for the TBR and the THR under different assumptions.

We estimated the "LE" bounds described in Section 3.1. Acute bronchitis is a kind of bronchial mucosal inflammation which is closely related to symptoms like fever, buccal thirst, throat itching, runny nose, dry stool, urine yellow, lung rale, tongue picture, and pulse condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

For each of these nine symptoms, we had a corresponding indicator covariate $\{X_i, i = 1, 2, ..., 9\}$ with $X_i \in \{0, 1\}$, where 1 stands for the presence of the corresponding symptom and 0 for not present. Let X_{LE} be the sum of these nine covariates, so $X_{\text{LE}} \in \{0, 1, 2, ..., 9\}$. The larger X_{LE}, the more serious acute bronchial the individual has. It is reasonable to assume that a individual with a relatively large X_{LE} would not be in the "always recover" group and that a individual with a relatively small X_{LE} would not be in the "never recover" group. We chose different sets for S_0 and S_1: for $0 \leq m_0 < m_1 \leq 9$, $S_1 = \{0, 1, ..., m_0\}$ and $S_0 = \{m_1, ..., 8, 9\}$.

The estimated "LE" bounds of the TBR and the THR under different values of m_0 and m_1 are shown in Table 1 in the supplementary materials. From the table, as m_0 increases or m_1 decreases, the bounds become narrower. This agrees with what we have seen previously, since larger values of m_0 and smaller values of m_1 lead to smaller values of $P(X \in S_2)$.

The individual is considered to have slight bronchitis if $X_{\text{LE}} \leq 1$ and very serious bronchitis if $X_{\text{LE}} = 9$. Thus, it seems reasonable to set $m_0 = 1$ and $m_1 = 9$. Under this setting, we have the estimated "LE" bounds for the TBR and the THR, as [0.498, 0.706] and [0.013, 0.221], respectively. The confidence intervals obtained by the bootstrap method are [0.358 0.797] and [0.000 0.306], respectively. It is also notable that the lower bound of the confidence interval for the TBR is larger than 0, which is a strong evidence that there exist at least 35.8% of individuals who can benefit from the treatment.

We used the "CI" method to estimate the TBR and the THR by assuming there exist at least three covariates that are independent conditional on $G = a, n$. The p-values of the J-test with the following combinations of symptoms are larger than 0.05:

1. (runny nose, dry stool, urine yellow, tongue picture),
2. (runny nose, dry stool, urine yellow, pulse condition),
3. (runny nose, dry stool, tongue Picture, pulse condition),

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

4. (runny nose, urine yellow, tongue Picture, pulse condition),

5. (dry stool, urine yellow, tongue Picture, pulse condition).

Only the fourth combination leads to a significant result for the TBR and the THR with the estimates 0.626 and 0.186, respectively, and the corresponding 95% confidence intervals (CI) [0.221, 1.000] and [0.000, 0.576], respectively. For the other combinations, the 95% CI for the TBR and the THR are all [0.00, 1.00], which may be due to the small sample size. The significant result shows a strong confidence that at least 22.1 percent of the population can benefit from the treatment.

The confidence intervals of the "LE" bounds are narrower than the "CI" estimators. This is consistent with the conclusion in the simulation study.

7 Discussion

Randomization is an effective tool to obtain the average causal effect of treatment versus control, but it is still important to assess the heterogeneity of treatment effects in the population. One way to characterize the treatment heterogeneity is to study the TBR and the THR. In this paper, we have proposed two methods for this. The "LE" bounds need covariates that can exclude the "always recover" stratum or the "never recover" stratum when the covariates belong to certain set; the "CI" assumption calls for at least three covariates that are independent in the "always recover" subgroup and the "never recover" subgroup.

For the "CI" estimators, we use more than three binary covariates $\{X_1, \cdots, X_k; k \geq 3\}$. When the observed covariate X_j is continuous or discrete with many values, we can dichotomize it by defining new covariate $\widetilde{X}_j = I(X_j > c_j)$ with a well-chosen constant c_j. Denote the optional set for c_j as C_j, $c = (c_1, \cdots, c_k), C = C_1 \times \cdots \times C_k$. Denote Σ_c as the covariance matrix of $(\widehat{\pi}_b, \widehat{\pi}_h)$ when X_j is dichotomized by truncating at $c_j, j = 1, \cdots, k$. We choose the optimal c by

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

minimizing the sum of the variances of $\widehat{\pi}_b$ and $\widehat{\pi}_h$, $c_{\text{opt}} = \arg\min_{c\in C} \text{tr}(\Sigma_c)$, where $\text{tr}(A)$ is the trace of A. In practice, to reduce the computation burden, we can choose C_j to be some sample quantiles of X_j.

In practice, we may collect various symptoms of the patients to satisfy Assumption 4. The suggestion is to use the covariate selection procedure described in Section 4.3 to choose appropriate covariates, and then estimate the TBR and the THR by the method proposed in Section 4.2. We can also estimate the "LE" bounds of the TBR and the THR and their confidence intervals by choosing S_0 and S_1. With the estimated confidence intervals of these two methods, we can choose the narrower intervals to get sharper inferences of the TBR and the THR.

The validation of Assumption 4 limits the use of the "CI" estimators. In many subgroup analyses, the covariates that define subgroups of patient populations are some biomarkers that often are not caused by the disease. It may not be reasonable to assume such covariates are independent given the latent principle strata unless the biomarkers of subgroups are some disease-caused factors. Nevertheless, usually there are many symptoms collected in a clinical trial which can be used as the possible covariates in Assumption 4.

Acknowledgements

We thank the Co-Editor and two reviewers for their insightful and constructive comments, which have greatly improved the manuscript. We also thank the Ministry of Science and Technology of the PRC for supporting research entitled "Significant New Drug Development– Construction of Technology Platform used for Original New Drug Research and Development"(2012ZX09303-010-002), which provided motivating data for this paper. Dr. Zhou's work was supported in part by U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, Health Science Research and Development grant (RCS 05-196).

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

Supplementary Materials

Refer to Web version on PubMed Central for supplementary materials.

References

[1] Albert, J. M., Gadbury, G. L., and Mascha, E. J. (2005). Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biometrical Journal* **47,** 662-673.

[2] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444-455.

[3] Berger, V. W., Rezvani, A., and Makarewicz, V. A. (2003). Direct effect on validity of response run-in selection in clinical trials. *Controlled Clinical Trials* **24,** 156-166.

[4] Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "Naive Bayes", and some alternatives when there are many more variables than observations, *Bernoulli* **10,** 989-1010.

[5] Davidoff, F. (2009). Heterogeneity is not always noise: Lessons from improvement. *The Journal of the American Medical Association* **302,** 2580-2586.

[6] Elrington, G. M., Murray, N. M., Spiro, S. G., and Newsom-Davis, J. (1991). Neurological paraneoplastic syndromes in patients with small cell lung cancer. A prospective survey of 150 patients. *Journal of Neurology, Neurosurgery Psychiatry* **54**, 764-767.

[7] Frangakis, C. E., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21-29.

[8] Gadbury, G. L. and Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics* **56,** 882-885.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

[9] Gadbury, G. L., Iyer, H. K., and Allison, D. B. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics* **11,** 313-333.

[10] Gadbury, G. L., Iyer, H. K., and Albert, J. M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference* **121,** 163-174.

[11] Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41,** 361-372.

[12] Goetghebeur, E. and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91,** 928-934.

[13] Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50,** 1029-1054.

[14] Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81,** 945-970.

[15] Long, D. M. and Hudgens, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics* **69,** 812-819.

[16] Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* **21,** 2917-2930.

[17] Poulson, R. S., Gadbury, G. L. and Allison, D. B. (2012). Treatment heterogeneity and individual qualitative interaction. *The American Statistician* **66,** 16-24.

[18] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41-55.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0174

[19] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688-701.

[20] Rubin, D. B. (1980). Comment on Randomization Analysis of Experimental Data: The Fisher Randomization Test by D. Basu, *Journal of the American Statistical Association* **75,** 591-593.

[21] Russek-Cohen, E. and Simon, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine* **16,** 455-464.

[22] Shen, C., Jeong, J., Li, X., Chen, P., and Buxton, A. (2013). Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect. *Biometrics* **69,** 724-731.

[23] Vermunt, J. K., Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis* **11,** 89-106.

[24] Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicineReporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357,** 2189-2194.

[25] Zhang, Z., Wang, C., Nie, L., and Soon, G. (2013). Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society, Serics C* **62,** 687-704.

[26] Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2012). Statistical methods in diagnostic medicine (Vol. **569**). *John Wiley & Sons*.