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

Notice: Accepted version subject to English editing.
ASSESSING THE HETEROGENEITY OF TREATMENT EFFECTS BY IDENTIFYING THE TREATMENT BENEFIT RATE AND TREATMENT HARM RATE

Yunjian Yin¹, Xiao-Hua Zhou¹²*, Zhi Geng¹**, and Fang Lu³

¹Peking University, ²University of Washington and ³China Academy of Chinese Medical Sciences

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), which are 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, respectively. We first 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. Then 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 (a) the proposed estimators work quite well when the conditional independence assumption hold, (b) the proposed estimators are not sensitive to small violation of the assumption, and (c) the proposed bounds we proposed can perform better than the estimators when the sample size is small. Finally, we
illustrate the application of the proposed methods in a real-world, 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. Therefore, the treatment effect can be heterogeneous, and 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 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 suffer from the treatment, i.e., 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), Poulsen, Gadbury, and Allison (2012)).

Some researchers made some extra assumptions regarding these proportions to resolve this heterogeneity, and one of the main assumptions is “Monotonicity” (Goetghhebeur and Molenberghs (1996) and Angrist, Imbens, and Rubin (1996)), which assumes that the treatment effect can not be worse than the control for every individual. However, 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 in this paper 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 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. Therefore we may want to obtain the bounds. Simple bounds for them can be easily derived without further assumptions. Gadbury, Iyer, and Albert (2004) tried to use the matched trial to obtain tighter bounds for the THR, but only found the same bounds as the simple bounds. Then they recognized that the THR could be identified in a well-matched design. However, the definition of a well-matched pair was based on the joint distribution of potential outcomes, which means it could not be measured in practice. As a result they tried to estimate the quality of matching in an extended matched design to derive tighter bounds. The extended matched design here means that two subjects in the same pair could both be assigned to the treatment or the control. However, it is possible that the bounds derived from this method could be worse than the simple bounds. Albert, Gadbury, and Mascha (2005) thought of the extended matched trial as a special case of the block trial, where each block contained subjects assigned to both treatment and control. With the assumption that subjects were exchangeable within blocks and that within-block probabilities were constant across blocks, they estimated the bounds and provided the variances for the estimators. Some researchers derived bounds conditional on covariates and then obtained the bounds in whole population by averaging over the distributions of covariates (e.g., Long and Hudgens (2013), Zhang, Wang, Nie, and Soon (2013)). Zhang, Wang, Nie, and Soon (2013) adjusted the simple bounds by covariates and they claimed that the adjusted bounds could not be worse than the simple bounds.

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 can not be tested by 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.

In this article, we propose new methods to study this problem from different angles. First, we try to use the covariates, which can provide some information for the joint distribution of the two potential outcomes, to obtain bounds for the TBR and the THR. The covariates can be either discrete, continuous, or multidimensional. The bounds we derive are always better than the adjusted simple bounds which were derived by Zhang, Wang, Nie, and Soon (2013). Then we show that we could identify the TBR and the THR under a different assumption, which does not need the conditional independence between two potential outcomes but 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 these estimators. Compared to the estimators derived by Shen, Jeong, Li, Chen, and Buxton (2013) and Zhang, Wang, Nie, and Soon (2013), our estimators have two merits. First, the crucial assumption for validation of our method can be tested by the observed data, while their assumption can not be tested. Second, in the estimation, we propose nonparametric estimators for the TBR and THR while their method requires parametric models.

We organize the paper as follows. We give the notation and assumptions in Section 2. In Section 3 we propose our better bounds for the TBR and the THR. Then we derive the nonparametric estimators in Section 4. The simulation results are reported in Section 5. We also apply our methods to real data analysis in Section 6. The paper ends up 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). 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 a univariate or a vector when needed. Before introducing some further notation, we need to make the following two assumptions, which are fundamental
and widely used in causal inference.

**Assumption 1.** (Stable unit treatment value assumption, (SUTVA)) There is no interference between units, which means that 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)).

The SUTVA assumption is reasonable in many cases. 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, which is defined as follows:

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

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.

With the notation above, Sheng, Jeong, Li, Chen, and Buxton (2013) defined the treatment benefit ratio (TBR) and the treatment harm ratio (THR) as follows:

\[
TBR : P(G = b) = P(Y_0 = 0, Y_1 = 1); \quad THR : P(G = h) = P(Y_0 = 1, Y_1 = 0).
\]

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

Under this assumption, we can easily identify the marginal distributions of \( Y_0 \) and \( Y_1 \), and then \( \text{ATE} = E[Y_1 - Y_0] \) can be identified. With a little calculation, we have the following equations:

\[
\begin{align*}
  P(G = b) + P(G = n) &= P(Y_0 = 0), \\
  P(G = h) + P(G = n) &= P(Y_1 = 0), \\
  P(G = a) + P(G = b) + P(G = h) + P(G = n) &= 1.
\end{align*}
\]

(1)

We have three equations and four parameters, so if one of the proportions of the four strata is identified or estimated, the others can also be easily identified or estimated.
3 Better bounds based on closely related covariates

As we mentioned in the introduction, the TBR and the THR cannot be identified even in randomized trials without further assumptions. We first try to derive bounds for these two rates. Denote \( p_1 = P(Y_1 = 1) \) and \( p_0 = P(Y_0 = 1) \), which can be easily identified in a randomized trial. Mathematically, it is easy to get the simple bounds for the TBR and the THR:

\[
\max(0, p_1 - p_0) \leq \text{TBR} \leq \min(1 - p_0, p_1), \quad \max(0, p_0 - p_1) \leq \text{THR} \leq \min(1 - p_1, p_0).
\]

These bounds, which are 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 \( P(Y_0 = 0) \) and \( P(Y_1 = 1) \). Similarly, the THR cannot be smaller than the negative ATE, which is equivalent to \( p_0 - p_1 \) 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. Specifically, let \( p_{1X} = P(Y_1 = 1 | X) \) and \( p_{0X} = P(Y_0 = 1 | X) \). The bounds for the TBR and the THR after being adjusted by \( X \) can be derived as follows:

\[
\begin{align*}
E[\max(0, p_{1X} - p_{0X})] & \leq \text{TBR} \leq E[\min(1 - p_{0X}, p_{1X})], \\
E[\max(0, p_{0X} - p_{1X})] & \leq \text{THR} \leq E[\min(1 - p_{1X}, p_{0X})].
\end{align*}
\]

(3)

Zhang, Wang, Nie, and Soon (2013) pointed out that the adjusted bounds, given by (3), can not be worse than the simple bounds given by (2).

The bounds in (3) only use the information about the relationship between \( X \) and \( Y_t, t = 0, 1 \). Sometimes, the covariates can also provide some information about the joint distribution of \( (Y_0, Y_1) \). We propose the following assumption, which can be used to further tighten the bounds.

Assumption 3. (Local Exclusion) Assume that \( S_0 \) and \( S_1 \) are two known subsets of the domain of \( X \) so that the subjects with \( X \in S_0 \) would not be in the “always recover” stratum, and the subjects with \( X \in S_1 \) would not be in the “never recover” stratum.

We call this assumption “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...
To interpret this assumption, we can 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. This assumption may be reasonable in some studies. Let us take drug therapy aiming at helping patients recover from bacterial infection inflammation as an example. Define $T = 1$ 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 while $Y = 0$ if not. Let $X$ be the indicator variable, representing the severity of inflammation with large values meaning severe inflammation. People with very serious inflammation can not recover from the placebo treatment, while people with mild inflammation can be easily cured by the drug treatment. Thus, the individual with large value of $X$ can not 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, it is easy to have that $P(G = a | X \in S_0) = P(G = n | X \in S_1) = 0$. From Equation (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. Specifically,

$$P(G = b | X \in S_0) = P(G \in \{a, b\} | X \in S_0) = P(Y_1 = 1 | X \in S_0) = P(Y = 1 | X \in S_0, T = 1),$$

$$P(G = b | X \in S_1) = P(G \in \{b, n\} | X \in S_1) = P(Y_0 = 0 | X \in S_1) = P(Y = 1 | X \in S_1, T = 0),$$

$$P(G = h | X \in S_0) = P(G \in \{a, h\} | X \in S_0) = P(Y_0 = 1 | X \in S_0) = P(Y = 1 | X \in S_0, T = 0),$$

$$P(G = h | X \in S_1) = P(G \in \{h, n\} | X \in S_1) = P(Y_1 = 0 | X \in S_1) = P(Y = 1 | X \in S_1, T = 1).$$

Let $S_2 = S_0 \cup S_1$, where $\cup$ stands for the complement operation. For the TBR and the THR in the subpopulation with $X \in S_2$, the bounds can be obtained by adjusting simple bounds with $X$ as in (3). Thus, we have the following theorem for the bounds of the TBR and the THR:

**Theorem 1.** Under Assumptions 1, 2 and 3, we can get the bounds of the TBR and the
THR as follows:

\[
\begin{align*}
\text{TBR} & \geq L_h = P(Y_1 = 1, X \in S_0) + P(Y_0 = 0, X \in S_1) + E[\max(0, p_1 - p_0 X) I(X \in S_2)], \\
\text{TBR} & \leq U_h = P(Y_1 = 1, X \in S_0) + P(Y_0 = 0, X \in S_1) + E[\min(p_1 X, 1 - p_0 X) I(X \in S_2)], \\
\text{THR} & \geq L_h = P(Y_0 = 1, X \in S_0) + P(Y_1 = 0, X \in S_1) + E[\max(0, p_0 X - p_1 X) I(X \in S_2)], \\
\text{THR} & \leq U_h = P(Y_0 = 1, X \in S_0) + P(Y_1 = 0, X \in S_1) + E[\min(p_0 X, 1 - p_1 X) I(X \in S_2)].
\end{align*}
\]

For simplicity, we denote these bounds as “LE” bounds (Local Exclusion). It is easy to see that (\(L_h, U_h\)) can be identified due to Assumption 2.

In addition, 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 \(P(X \in S_2) = 0\).

Also, the following proposition states an obvious result that 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 “LE” bounds while the bounds in (3) do not. We leave the detailed proof in the supplementary material.

**Proposition 1.** The “LE” bounds for the TBR and the THR we derived can not be worse than the bounds in (3), and 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 of the TBR and the THR can be easily estimated by the moment estimator. Denote \(\hat{L}_h, \hat{U}_h, \tilde{L}_h, \tilde{U}_h\) as the resulting non-parametric estimators for the lower and upper bounds of the TBR and the THR, respectively. These non-parametric estimators have the following forms:

\[
\begin{align*}
\hat{L}_h &= \frac{P_n[f(1,0,0)]}{P_n[I(T = 0)]} + \frac{P_n[f(0,1,1)]}{P_n[I(T = 1)]} + \sum_{s \in S_2} \max\left\{0, \frac{P_n[g(1,x,0)]}{P_n[I(T = 0)]}, \frac{P_n[g(1,x,1)]}{P_n[I(T = 1)]}\right\}, \\
\hat{U}_h &= \frac{P_n[f(0,0,0)]}{P_n[I(T = 0)]} + \frac{P_n[f(0,1,1)]}{P_n[I(T = 1)]} + \sum_{s \in S_2} \min\left\{0, \frac{P_n[g(1,x,0)]}{P_n[I(T = 0)]}, \frac{P_n[g(0,x,1)]}{P_n[I(T = 1)]}\right\}, \\
\tilde{L}_h &= \frac{P_n[f(1,0,1)]}{P_n[I(T = 1)]} + \frac{P_n[f(0,1,0)]}{P_n[I(T = 0)]} + \sum_{s \in S_2} \max\left\{0, \frac{P_n[g(1,x,1)]}{P_n[I(T = 1)]}, \frac{P_n[g(1,x,0)]}{P_n[I(T = 0)]}\right\}, \\
\tilde{U}_h &= \frac{P_n[f(0,0,1)]}{P_n[I(T = 1)]} + \frac{P_n[f(0,1,0)]}{P_n[I(T = 0)]} + \sum_{s \in S_2} \min\left\{0, \frac{P_n[g(0,x,0)]}{P_n[I(T = 0)]}, \frac{P_n[g(1,x,1)]}{P_n[I(T = 1)]}\right\},
\end{align*}
\]

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 the confidence intervals for the lower bounds and upper bounds. Specifically, we first randomly draw datasets from the original sample with replacement. With the new dataset, we can compute estimates of \((L_b, U_b, L_h, U_h)\), denoted as \((\hat{L}_b, \hat{U}_b, \hat{L}_h, \hat{U}_h)\). This process is done \(B\) times, so we have \(B\) bootstrap replications, \(\left(\hat{L}_b, \hat{U}_b, \hat{L}_h, \hat{U}_h\right)\), \(\cdots\), \(\left(\hat{L}_b, \hat{U}_b, \hat{L}_h, \hat{U}_h\right)\). We can form approximate 95% confidence intervals by finding the 2.5% and 97.5% percentiles of \((\hat{L}_b, \hat{U}_b, \hat{L}_h, \hat{U}_h)\), denoted as \(\left(\hat{L}_{b,(2.5)}^*, \hat{U}_{b,(2.5)}^*, \hat{L}_{h,(2.5)}^*, \hat{U}_{h,(2.5)}^*\right)\) and \(\left(\hat{L}_{b,(97.5)}^*, \hat{U}_{b,(97.5)}^*, \hat{L}_{h,(97.5)}^*, \hat{U}_{h,(97.5)}^*\right)\), respectively. Then the approximate 95% confidence intervals for the bounds of the TBR and the THR can be constructed as:

\[
[\hat{L}_{b,(2.5)}^*, \hat{U}_{b,(97.5)}^*], \quad [\hat{L}_{h,(2.5)}^*, \hat{U}_{h,(97.5)}^*].
\]

4 Nonparametric Identifiability and Estimation

Zhang, Wang, Nie, and Soon (2013) and Shen, Jeong, Li, Chen, and Buxton (2013) proposed the conditionally independence assumption to identify the TBR and the THR in a randomized trial: \(Y_1 \perp Y_0 | X\). Under this assumption, the joint distribution of \((Y_0, Y_1)\) can be identified by factorization, i.e., \(P(Y_0, Y_1|X) = P(Y_0|X)P(Y_1|X)\). However, in practice it is hard to find the covariates \(X\) that can explain all the dependence between \(Y_0\) and \(Y_1\). Even if such \(X\) is observed, \(X\) is usually high-dimensional, so the estimation is still a complex and difficult task. They considered estimating the TBR and the THR by assuming parametric models between \(X\) and potential outcomes, which may lead to large bias if the models are misspecified.

In this Section, we first consider the nonparametric identification of the TBR and the THR under a different 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 \(X = (X_1, \ldots, X_k)\). We assume that \(X_j, j = 1, \ldots, k\), are binary variables; this binary assumption is just for convenience in the derivation of identification and is not necessary because identification can be easily achieved for non-binary covariates by dichotomizing them. The dichotomization will be discussed in the discussion section.

Assumption 4. \(X_1, \ldots, X_k\) are mutually independent in the “always recover” group and
“never recover” group, i.e., $X_1, \ldots, X_k$ are conditionally independent given $G = a, n$.

Assumption 4 can be true with some properly chosen $X_1, \ldots, X_k$ in some settings. Let us consider treatment as a medicine or a therapy aimed at curing a certain kind of disease, define the outcome as a binary variable, indicating whether a patient is cured (1 if cured and 0 if not), and 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 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. Thus 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. Our outcome is whether the patient is cured at the end of the trial. Let $X_1, \ldots, X_k$ be the diagnosis on the severity of the disease by $k$ different doctor with different medical background. Therefore, given the true but latent severe level of the disease, $X_1, \ldots, X_k$ are most likely conditional independent because $k$ doctors made their own diagnoses on the severity of disease, based on their own experiences rather than any other common principles or 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 $X_1, \ldots, X_k$ are independent conditional on $G = a, n$. Therefore, Assumption 4 holds for the chosen variables, $X_1, \ldots, X_k$.

In practice, this kind of assumption has been used in many other settings. It is well known that the naive Bayes classifier uses the assumption that the covariates are mutually independent given the true class and 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)). Also, 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, which is highly related to \( G = n \) and \( G = a \) groups, then Assumption 4 most likely holds for these variables.

Under Assumption 4, we can have the following probability decomposition:

\[
P(X_1,\ldots,X_k, Y = 1| T = 1) = P(X_1,\ldots,X_k| G = a)\pi_a + P(X_1,\ldots,X_k| G = b,T = 1, Y = 1)\pi_b
\]

\[
= P(X_1| G = a)\cdots P(X_k| G = a)\pi_a + P(X_1,\ldots,X_k| G = b)\pi_b,
\]

where \( \pi_g = p(G = g) \), \( g = a, b, h, n \). Similarly, we have the following result:

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

What we want to identify is \( \pi = (\pi_a, \pi_b, \pi_h, \pi_n) \). By rearranging the equations above to eliminate some nuisance parameters, we can have the following result:

\[
P(X_1,\cdots,X_k, Y = 1| T = 1) - P(X_1,\cdots,X_k, Y = 0| T = 0)
\]

\[
= P(X_1| G = a)\cdots P(X_k| G = a)\pi_a - P(X_1| G = n)\cdots P(X_k| G = n)\pi_n.
\]

We have \( 2^k \) equations, and \( 2(k+1) \) parameters here, hence to identify the parameters we need that: \( 2^k \geq 2(k+1) \), i.e., \( k \geq 3 \). But having more equations than parameters is not enough to guarantee an unique solution, we still need the following assumption:

**Assumption 5.** There exists at least one covariate in \( \{X_1,\ldots,X_k\} \), say \( X_j \), such that

\( P(X_j| G = a) \neq P(X_j| G = n) \).

This assumption means that the distribution of at least one component of \( X \) conditional on \( G = a \) should not be the same with that conditional on \( G = n \). With this assumption, we can have the following theorem, with a proof given in the supplementary material.

**Theorem 2.** When \( k \geq 3 \), under Assumptions 1, 2, 4 and 5, the TBR and the THR are identifiable.

Theorem 2 shows that if there are at least three covariates \( \{X_1,\cdots,X_k\} \), which are independent conditional on \( G = a \) and \( G = n \), then the TBR and the THR can be identifiable when Assumption 5 holds true.
4.2 Nonparametric estimation

The nonparametric estimators proposed in this section are based on the generalized method of moments (GMM) estimator, which was formalized by Hansen (1982). For simplicity of introducing our idea, we assume all $X_1, \ldots, X_k$ are binary variables. The non-binary case will be discussed in the discussion section.

Denote $\rho_{a} = P(X_j = 1|G = a), \rho_{n_j} = P(X_j = 1|G = n), \pi_g = P(G = g), g = a, b, h, n, p_1 = P(T = 1), p_2 = P(Y = 1|T = 1), \theta = \{\pi_b, \pi_h, \rho_{a1}, \ldots, \rho_{nk}, \pi_n, p_1, p_2\}$. By substituting $\pi_a = p_2 - \pi_b, \pi_n = 1 - p_2 - \pi_h$ into Equation (4), we can have the following result:

$$P(X_1 = x_1, \ldots, X_k = x_k, Y = 1|T = 1) - P(X_1 = x_1, \ldots, X_k = x_k, Y = 0|T = 0) = (p_2 - \pi_h)\varphi_{a1}(x_1) \cdots \varphi_{nk}(x_k) - (1 - p_2 - \pi_h)\varphi_{n1}(x_1) \cdots \varphi_{nk}(x_k),$$

where $\varphi_{ij}(x_j) = \rho_{ij}^{x_j}(1 - \rho_{ij})^{1-x_j}, g = a, n, j = 1, \ldots, k$. Denote

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

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

$$g(x_1, \ldots, x_k; \theta) = P_n[\hat{g}(x_1, \ldots, x_k; \theta)]$$

$$= P_n\left[I(X_1 = x_1, \ldots, X_k = x_k, Y = 1, T = 1)/p_1 - I(X_1 = x_1, \ldots, X_k = x_k, Y = 0, T = 0)/(1 - p_1)\right]$$

$$- (p_2 - \pi_h)\varphi_{a1}(x_1) \cdots \varphi_{nk}(x_k) - (1 - p_2 - \pi_h)\varphi_{n1}(x_1) \cdots \varphi_{nk}(x_k)],$$

and

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

where $I(\cdot)$ is the indicator function. Then the GMM estimator $\hat{\theta}_n$ can be defined as follows:

$$\hat{\theta}_n = \arg\min_\theta Q(\theta) = \arg\min_\theta g(\theta)^TW(\theta)^{-1}g(\theta),$$

(5)

where $W(\theta)$ is a positive semi-definite matrix. To reduce the computational burden, we use the
two-step procedure to estimate \( \theta \). Specifically, the first step estimates are constructed by using a preliminary weighting matrix \( \hat{W} \) (the identify matrix is used here), to replace \( W(\theta) \) in (5), and let \( \hat{\theta}_{n,1} \) be a solution to the initial optimization problem,

\[
G(\hat{\theta}_{n,1})^T \hat{W}^{-1} g(\hat{\theta}_{n,1}) = 0,
\]

where \( G(\theta) = \frac{\partial}{\partial \theta} g(\theta) \). Denote \( S(\theta) \) as the sample covariance matrix of \( g(\theta) \). Then the two-step estimator \( \hat{\theta}_n \) is defined as in (5) by replacing \( W(\theta) \) with \( S(\hat{\theta}_{n,1}) \). Specifically, \( \hat{\theta}_n \) solves the following estimating equation:

\[
G(\hat{\theta}_n)^T S(\hat{\theta}_{n,1})^{-1} g(\hat{\theta}_n) = 0.
\]

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

**Theorem 3.** Under Assumptions 1, 2, 4, 5 with \( k \geq 3 \), the asymptotic distribution of \( \hat{\theta}_n \) is

\[
\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, (G^T S^{-1} G)^{-1}),
\]

where \( G = \lim_n G(\theta), S = \lim_n S(\theta) \). Particularly, the estimators \( (\hat{\pi}_b, \hat{\pi}_h) \) of \( (\pi_b, \pi_h) \) has the following asymptotic distribution:

\[
\sqrt{n}( \begin{pmatrix} \hat{\pi}_b \\ \hat{\pi}_h \end{pmatrix} - \begin{pmatrix} \pi_b \\ \pi_h \end{pmatrix} ) \xrightarrow{d} N(0, \Sigma),
\]

where \( \xrightarrow{d} \) means convergence in distribution, \( \Sigma \) is the corresponding \( 2 \times 2 \) block matrix of \( (G^T S^{-1} G)^{-1} \).

For simplicity, we denote \( (\hat{\pi}_b, \hat{\pi}_h) \) as “CI” estimators (conditionally independent) and Assumption 4 with \( k \geq 3 \) as “CI” assumption.

The variance \( (G^T S^{-1} G)^{-1} \) can be estimated by \( (G^T (\hat{\theta}_n) S(\hat{\theta}_{n,1})^{-1} G(\hat{\theta}_n))^{-1} \). Thus we can also have an estimator for \( \Sigma \), denoted by \( \hat{\Sigma} \). Then the 95% confidence intervals for \( \pi_b \) and \( \pi_h \) can be constructed as

\[
[\hat{\pi}_b - 1.96\sqrt{\hat{\Sigma}_{11}}, \hat{\pi}_b + 1.96\sqrt{\hat{\Sigma}_{11}}], \; [\hat{\pi}_h - 1.96\sqrt{\hat{\Sigma}_{22}}, \hat{\pi}_h + 1.96\sqrt{\hat{\Sigma}_{22}}],
\]

where \( \hat{\Sigma}_{ij} \) is the corresponding element in the matrix \( \hat{\Sigma} \).
4.3 Over-identified test and backward variables selection

In practice, people may question the validation of Assumption 4. The GMM method provides a test called over-identified test or J-test when we have more than three covariates. Specifically, we define the J-statistics as follows:

$$J = ng(\hat{\theta}_n)^T W(\hat{\theta}_n)^{-1} g(\hat{\theta}_n) \rightarrow \chi^2(2^k - 2k - 2).$$

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 J-test to select appropriate symptoms. The covariate selection procedure is described as follows:

1. Initialize $X_{\text{new}} = X$.
2. Calculate the J-test statistics and its p-value with the covariates $X_{\text{new}}$, if the p-value is bigger than 0.05 or $\text{dim}(X_{\text{new}}) \leq 4$, then stop; if not, go to step 3.
3. Remove the $r$-th component $X_{\text{new},r}$ from $X_{\text{new}}$, saying $X_{\text{new},-r}$, where $r = 1, \cdots, \text{dim}(x_{\text{new}})$, and calculate the corresponding J-statistics $J_r$ and p-value $p_r$; and update

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

Then go back to step 2.

If the final 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 conducted. First, we evaluate the performances of the “CI” estimators when the “CI” assumption holds and does not hold, respectively. The performance was measured by bias, and bias percentage, which was defined by $100 \times \frac{\text{bias/true value}}{\%}$, i.e., the absolute value of bias over true value. We also estimated the
ASSESSING THE HETEROGENEITY OF TREATMENT EFFECTS

average asymptotic standard error (ASE), the empirical standard error (ESE) and the coverage of 95% confidence intervals. Second, we compare the average length of the confidence intervals (ALCs) 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, ξ, ξ_k, k = 1, 2, 3, 4), which would be used later. 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 \{P(G = g), g = a, b, h, n\} = (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. Once we had those six variables, then we constructed the covariates as follows:

\[
\begin{align*}
\text{In the subgroup } G = g, \text{ let } \bar{X}_k &= \mu_g + \alpha_{g,k}\xi + \xi_k, \quad k = 1, 2, 3, 4, \quad g = a, b, h, n, \\
X_k &= I(\bar{X}_k > 0).
\end{align*}
\]

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 \(\alpha_a = (\alpha_{a,1}, \alpha_{a,2}, \alpha_{a,3}, \alpha_{a,4}) = \alpha_n = (\alpha_{n,1}, \alpha_{n,2}, \alpha_{n,3}, \alpha_{n,4}) = \gamma \cdot (1, 1, 1, 1)\). It is easy to see that the “CI” assumption holds when \(\gamma = 0\). As \(\gamma\) increases from 0, the correlation between \((X_1, \ldots, X_4)\) conditional on \(G = a, n\) increases. The larger value of \(\gamma\) can cause large violation of Assumption 4, which may induce large bias of the “CI” estimators.

With each data set, we first 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 of which the p-values are 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, i.e., \(\gamma = 0\), the estimators perform very well with small bias and bias percentage. The reject 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 the \(\gamma\) 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%, this indicates that our estimators can still perform very well. Thus, our estimators are not sensitive to small violation of the assumption. When \(\gamma\) continues to increase (\(\gamma > 0.5\)), the bias and the bias percentage do not increase a lot. Also note that the
Table 1: The performance of our estimators, J-test and the coverage of 95% confidence intervals

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

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.
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 coverage of the 95% confidence intervals decreases rapidly, which is only nearly 70% when $\gamma = 1$.

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

Step 1: A set of values for the sample size $n$ was created. Variables $T$ and $G$ were generated independently. More specifically, the treatment $T$ was generated from a Bernoulli distribution with $P(T = 1) = 0.5$, the principal strata $G$ was generated by randomly drawing from $\{a, b, h, n\}$ with probabilities $(0.4, 0.3, 0.2, 0.1)$. With $T$ and $G$ being generated, the outcome $Y$ could be decided by the definition of $G$. The covariate $X_{LE}$ that was used in obtaining the “LE” bounds and the covariates $X_{CI} = (X_{CI,1}, \ldots, X_{CI,4})$ that was used in “CI” estimators were generated independently given $G$. The distribution of $X_{LE}$ was as follows:

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

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

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

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

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

In the subgroups $G = b, h$, we first randomly generated $(\epsilon, \epsilon_1, \ldots, \epsilon_4)$, which were mutually independent and standard normally distributed. Then the covariates $X_{CI}$ were constructed as follows:

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

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

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

Step 3: Step 1 and Step 2 were repeated 1000 times to estimate the ALCIs of “LE” bounds and “CI” estimators, respectively.

Figure 1 shows the ALCIs for the TBR and the THR under different sample sizes. From the figure 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 speed of 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.

6 Analysis of a randomized clinical trial

We now apply 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 little 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. This means that the treatment has a better average effect than the placebo. But there may still exist individuals who are harmed by the treatment.
Figure 1: The comparison of the length of the confidence intervals for CI estimators and LE bounds.
We applied the methods proposed in this paper to obtain the bounds and estimations for the TBR and the THR under different assumptions, respectively.

We first estimated the “LE” bounds, which was 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. For each of these nine symptoms, we have a corresponding indicator covariate, and denote the nine indicator variables by \( \{X_i, i = 1, 2, ..., 9\} \). Here \( X_i \in \{0, 1\} \), 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 \( X_{LE} \in \{0, 1, 2, ..., 9\} \). The larger \( X_{LE} \) is, the more serious acute the individual has. So 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. Based on this reason, we choose different sets for \( S_0 \) and \( S_1 \). Specifically, let \( 0 \leq m_0 < m_1 \leq 9 \), and define \( S_1 = \{0, 1, ..., m_0\} \) 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 material. We can see from the table that as \( m_0 \) increases or \( m_1 \) decreases, the bounds become narrower. This agrees with what we have obtained 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 \( X_{LE} \leq 1 \) and very serious bronchitis if \( X_{LE} = 9 \). Thus, it is 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.

Next, we used the “CI” method to estimate the TBR and the THR by assuming there exist at least three covariates, which 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),
4. (runny nose, urine yellow, tongue Picture, pulse condition),

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

However, only the fourth combination leads to significant result for the TBR and the THR with the estimates being 0.626 and 0.186, respectively, and the corresponding 95% confidence intervals (CI) being [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.

Note that the confidence intervals of the “LE” bounds are narrower than the “CI” estimators. This is consistent with the conclusion in the simulation study that the “LE” bounds have narrower confidence intervals when the sample size is small.

7 Discussion

Randomization is an effective tool to obtain the average causal effect of treatment versus control. However, it is still important to assess the heterogeneity of treatment effects in the population, and one way to characterize the treatment heterogeneity is to study the TBR and the THR. In this paper, we have proposed two methods to study the TBR and the THR, including the “LE” bounds and the estimators for the TBR and the THR. The “LE” bounds need covariates that can exclude the “always recover” stratum or “never recover” stratum when the covariates belong to certain set. The “CI” assumption enables us to identify and estimate the TBR and the THB, which calls for at least 3 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, \ldots, 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 \(\tilde{X}_j = I(X_j > c_j)\) with a well-chosen constant \(c_j\). Denote the optional set for \(c_j\) as \(C_j = (c_{j1}, \ldots, c_{jk}), C = C_1 \times \cdots \times C_k\). Denote \(\Sigma_e\) as the covariance matrix of \((\hat{\pi}_b, \hat{\pi}_h)\) when \(X_j\) is dichotomized by truncating at \(c_j, j = 1, \ldots, k\). We choose the optimal \(c\) by minimizing the sum of the variances of \(\hat{\pi}_b\) and \(\hat{\pi}_h\), which is the trace of \(\Sigma_e\); specifically,

\[c_{opt} = \arg \min_{c \in C} \text{tr}(\Sigma_e),\]
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 $X_j$. For example, we can choose $C_j = (\rho_{0.2}(X_j), \rho_{0.4}(X_j), \rho_{0.6}(X_j), \rho_{0.8}(X_j))$, where $\rho_{\tau}(X_j)$ is the sample $\tau$th quantile 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 a sharper inference of the TBR and the THR.

Note that the validation of Assumption 4 limits the use of the “CI” estimators in practice. In many subgroup analyses, the covariates which 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 and thus our proposed method can still be widely used.

**Acknowledgements**

We thank the Co-Editor and two reviewers for their insightful and constructive comments, which have greatly improved the manuscript. We also wish to thank the Ministry of Science and Technology of the PRC for supporting a 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).

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.
References


