Statistica Sinica Preprint No: SS-2016-0150R1	
Title	Assessing The Treatment Effect Heterogeity with a Latent Variable
Manuscript ID	SS-2016-0150R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0150
Complete List of Authors	Yunjian Yin Lan Liu and Zhi Geng
Corresponding Author	Lan Liu
E-mail	liux3771@umn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

ASSESSING THE TREATMENT EFFECT HETEROGENEITY WITH A LATENT VARIABLE

Yunjian Yin[1,2], Lan Liu[2,*], and Zhi Geng[1]

[1]*School of Mathematical Sciences, Peking University, Beijing 100871, China*

[2]*School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA*

Yunjian Yin: yinyunjian@pku.edu.cn; Lan Liu: liux3771@umn.edu; Zhi Geng: zhigeng@pku.edu.cn

[*]*Corresponding author phone: 1-919-593-8505; fax: 612-624-8868*

Abstract: The average treatment effect (ATE) is commonly used to assess the effect of treatment. However, the ATE implicitly assumes a homogenous treatment effect even amongst individuals with different characteristics. In order to describe the magnitude of heterogeneity, we define the treatment benefit rate (TBR) as the proportion of individuals in different subgroups who benefit from the treatment and define the treatment harm rate (THR) as the proportion harmed. These rates involve the joint distribution of the potential outcomes and cannot be identified without further assumptions, even in randomized clinical trials. Under the assumption that the potential outcomes are independent conditional on the observed covariates and an unmeasured latent variable, we show the identification of the TBR and THR in non-separable (generalized) linear mixed models for both continuous and binary outcomes. We then propose estimators and derive their asymptotic distributions. The proposed meth-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

ods are implemented in an extensive simulation study and two randomized controlled trials.

Key words and phrases: Average treatment effect; Causal inference; Heterogeneity; Latent Variable.

1 Introduction

The average treatment effect (ATE) is used in evaluating the effect of a treatment or an intervention in a wide range of disciplines such as medicine, social sciences, econometrics, etc. An assumption implicitly made by the ATE is the similarity of treatment effect across heterogeneous individuals. Although this assumption is warranted for some treatments, it is less plausible for others. For example, most patients treated with MMR (measles, mumps, and rubella) vaccine benefit from a very low risk of having Measles (one dose of MMR vaccine is about 93% effective while two doses are about 97% effective at preventing measles if exposed to the virus). In contrast, clinical evidence was found that prescription of a beta-blocker may or may not provide the desired response in treating patients with hypertension (Bradley, Mayosi, Maroney, Mbewu, Opie, and Volmink (2007)). Likewise, the prescription of anti-anxiety drugs such as Benzodiazepines may or may not be effective in treating patients with anxiety: some patients suffer from side effects such as drowsiness and depression while some others experienced paradoxical reactions such as increased anxiety, irritability, and agitation.

Formally, the heterogeneity of treatment effect is present if the effect of the treatment varies across subsets of individuals in a population (Poulson, Gadbury, and Allison (2012)). This variability at the individual level is also called subject-treatment interaction (Gadbury, Iyer, and Allison (2001); Gadbury, Iyer, and Albert (2001)). The heterogeneity of treatment effect may

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

not only arise from different baseline characteristics of individuals (also known as pre-treatment heterogeneity) such as age, sex and social status but also from distinct individual responses to a particular treatment or intervention (Brand and Thomas (2013)).

From a clinician's perspective, the heterogeneity of treatment effect plays an essential role in selecting the most effective treatment and designing individualized treatment regimens (Imai and Ratkovic (2013)). Specifically, a treatment with large response variability among patients should be used with more vigilant than a treatment with similar ATE but smaller variability. From a pharmaceutical company's perspective, it is crucial to identify and target the individuals that would benefit from the treatment. Finally, it is also critical for policy makers to understand treatment effect heterogeneity so as to generalize causal effect estimates obtained from an experimental sample to a target population.

With observed effect modifiers, the conditional ATE for different subpopulations is typically calculated. In principle, such subgroup analysis would yield homogeneous treatment effect controlling for all effect modifications. However, it is hard to target and collect all effect modifiers based on the existing knowledge and limited resources (Zhang et al. (2013)). As a result, the residual heterogeneity stand in the way of better understanding the treatment effect and more effectively designing the optimal treatment for each individual. Furthermore, the evidence of heterogeneous treatment effect urges further pursuit of unknown effect modifiers. Novel methods are thus of demand to assess the treatment effect heterogeneity of the study population or subpopulation.

To better illustrate the treatment effect heterogeneity, we use the framework of potential outcomes (Rubin (1974); Rosenbaum and Rubin (1983); Holland (1986)). Under this framework, each individual has a potential outcome for every possible treatment, and the individual level effect of an experimental treatment relative to a control is defined by a comparison between the corresponding potential outcomes. However, for each individual, only one potential outcome, the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

one corresponds to the actual treatment, can be observed in practice.

Under the potential outcomes framework, the treatment benefit rate (TBR) and the treatment harm rate (THR) have been defined to assess the treatment effect heterogeneity (Gadbury and Iyer (2000); Gadbury, Iyer, and Allison (2001); Gadbury, Iyer, and Albert (2004); Albert, Gadbury, and Mascha (2005); Poulson, Gadbury, and Allison, (2012); Shen, Jeong, Li, Chen, and Buxton (2013); Zhang, Wang, Nie, and Soon (2013)). When the outcomes are binary, the TBR (THR) is the proportion of individuals in different subgroups who have better (worse) outcomes if given the treatment compared with control. We define the TBR and THR similarly for continuous outcomes by comparing the difference between the potential outcomes with some level c. The definitions of the TBR and THR involve the joint distribution of the two potential outcomes, thus cannot be identified without further assumptions even in randomized trials.

Various bounds have been derived for the TBR and THR. Gadbury, Iyer, and Albert (2004) derived the simple bounds of the THR by using only the observed data without further assumptions, along with tighter bounds by estimating the quality of matching in a matched design. Albert, Gadbury, and Mascha (2005) extended the results to block trials that include the matched trial as a special case. **?** used a secondary outcome to obtain tighter bounds under monotonicity, transitivity and causal necessity assumptions.

To identify and estimate the TBR and THR, Shen, Jeong, Li, Chen, and Buxton (2013) and Zhang, Wang, Nie, and Soon (2013) assumed that the two potential outcomes were independent conditional on observed covariates. However, this assumption is stringent in practice since the two potential outcomes are from the same individual and there is no guarantee that all the observed covariates are sufficient to explain the dependence. Yin, Zhou, Geng, and Lu (2016) estimated the TBR and THR assuming the existence of at least three covariates which are mutually independent. Their assumption could be tested when more than three such covariates were available without

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

any modeling assumptions. However, it is hard to find such covariates in practice.

In this article, we make the weaker assumption that the potential outcomes are independent given the observed covariates and an unmeasured latent variable. Under non-separable (generalized) linear mixed models, we prove identification and construct estimators using maximum-likelihood estimation (MLE). All parameters in the models can be identified, including the coefficients corresponding to the unmeasured latent variable. Thus, we can empirically test whether it is necessary to include such unmeasured latent variable in the independence assumption and whether our models are indeed non-seperable. Moreover, we derive the asymptotic distributions and variances for the estimators.

We organize the paper as follows. In Section 2, we introduce the notation and describe the assumptions. In Section 3, we provide identification conditions for the TBR and THR under non-separable models for continuous and binary outcomes, respectively. The estimators and their asymptotic properties are derived in Section 4. We report simulation results in Section 5. We illustrate our proposed method in two randomized trials in Section 6. The paper concludes with a discussion in Section 7.

2 Preliminaries

Let T denote a binary treatment assignment variable that is completely randomized and let Y denote a primary outcome of interest. Let $X = (X_1, \cdots, X_p)^T$ denote a p–dimensional observed covariate, where the superscript ^{T} denotes transposition. Let t denote a possible value T could take ($t = 1$ for treatment and $t = 0$ for placebo). Assume a larger value of Y indicates better response. Under the Stable Unite Treatment Value Assumption (SUTVA) (**?**), let Y_1 and Y_0 denote the potential outcomes under treatment and control, respectively.

When the outcome variable Y is binary, define the TBR (THR) for the subpopulation with

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

specific covariate value $X = x$ as

$$\text{TBR}(x) = P(Y_0 = 0, Y_1 = 1|X = x) \text{ and } \text{THR}(x) = P(Y_0 = 1, Y_1 = 0|X = x)^*.$$

The TBR(x) is the proportion of individuals in the subpopulation with covariates $X = x$ that have better outcomes if given treatment compared to control. In contrast, the THR(x) is the proportion of individuals in the subpopulation that have better outcomes if given control compared to treatment. Let $\text{ATE}(x) = E(Y_1 - Y_0|X = x)$ denote the average treatment effect among subgroup with covariates $X = x$. When the outcomes are binary, we have $\text{ATE}(x) = \text{TBR}(x) - \text{THR}(x)$, that is, the subgroup ATE is the difference between the beneficial and harmful rates of the subgroup. Thus, TBR(x) and THR(x) not only provide information about the overall treatment effect but also how treatment effect may vary across individuals.

When Y is continuous, we extend the definition of the TBR(x) and the THR(x) by comparing the difference between the potential outcomes with some level c:

$$\text{TBR}_c(x) = P(Y_1 - Y_0 > c|X = x) \text{ and } \text{THR}_c(x) = P(Y_0 - Y_1 > c|X = x),$$

where c is a pre-specified constant. The $\text{TBR}_c(x)$ is the proportion of individuals in the subpopulation whose outcome Y would benefit greater than c from the treatment compared with the control and $\text{THR}_c(x)$ is the proportion of the individuals whose outcome Y would be harmed by at least c by the treatment compared with the control. A special case is when $c = 0$, $\text{TBR}_{c=0}(x)$ ($\text{THR}_{c=0}(x)$) denotes the proportion of the individuals whose outcomes benefit from (harmed by) the treatment regardless of the magnitude. If $Y_1 - Y_0$ is a continuous variable, we have $\text{TBR}_c(x) = 1 - P(Y_1 - Y_0 < c|X = x) = 1 - P(Y_0 - Y_1 > -c|X = x) = 1 - \text{THR}_{-c}(x)$. Thus,

* When the outcome is binary, a similar definition for the TBR and THR on the *population* level was proposed by Shen, Jeong, Li, Chen, and Buxton (2013). Throughout the paper, we focus on the *subpopulation* TBR(x) and THR(x) since it provides more detailed information on how the heterogeneity changes across different subgroups.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

we only consider the case of $c \geq 0$. When $c = 0$, we have $\text{TBR}_{c=0}(x) = 1 - \text{THR}_{c=0}(x)$, that is, regardless of the magnitude, the proportions of individuals who would benefit and be harmed sum to 1. This does not hold if $Y_1 - Y_0$ has a point mass at 0.

From the definitions of $\text{TBR}_c(x)$ and $\text{THR}_c(x)$, $1 - \text{TBR}_c(x)$ or equivalently $\text{THR}_{-c}(x)$, is the cumulative distribution function of $Y_1 - Y_0$ conditional on X. That is, $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ characterize the entire distribution of treatment effect rather than the mean in the subpopulation. Hence, the heterogeneity of treatment effect can be inferred from the $\text{TBR}_c(x)$ and $\text{THR}_c(x)$. Similar to the binary case, we can obtain the $\text{ATE}(x)$ from the $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ as $\text{ATE}(x) = \int_0^\infty \{\text{TBR}_c(x) - \text{THR}_c(x)\} dc$ (Appendix A). However, one cannot fully recover $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ from $\text{ATE}(x)$. Thus, $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ provide more information on the subgroup treatment effect than does $\text{ATE}(x)$.

Due to the randomization, we can identify the marginal distributions of Y_0 and Y_1 (conditional distributions of Y_0 given X and that of Y_1 given X) as well as $\text{ATE}(x)$. However, $\text{TBR}(x)$ and $\text{THR}(x)$, $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ involve the joint distribution of the two potential outcomes, thus cannot be identified even in randomized trails without further assumptions. To make progress, Shen, Jeong, Li, Chen, and Buxton (2103) and Zhang, Wang, Nie, and Soon (2013) made the following assumption, where $\perp$ denotes independence between variables.

Assumption 1. (Conditional Independence) $Y_0 \perp Y_1 | X$.

Assumption 1 states that the two potential outcomes are independent conditional on a set of observed baseline covariates. Hence, the joint distribution of Y_0 and Y_1 can be identified by the factorization $P(Y_0, Y_1|X) = P(Y_0|X)P(Y_1|X)$. However, this assumption requires the collection of relevant covariates X to control for all the dependency between two potential outcomes, which is hard to satisfy in practice and impossible to test from the observed data. Alternatively, we make an assumption that there is independence between the potential outcomes conditional on

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

observed covariates X and a latent variable U.

Assumption 2. (Latent Independence) $Y_0 \perp Y_1 | (X, U), \quad U \perp X$.

Assumption 1 is a special case of Assumption 2 when there is no latent variable U. The independence between X and U can be relaxed to a decomposition of U into any function of X and a random error ϵ_u that is independent of X (Appendix G). For the ease of illustration, we assume U and X are independent. Zhang, Wang, Nie, and Soon (2013) claimed that, under Assumption 2, the information of U is not identifiable in a generalized linear mixed model (GLMM) and thus adopted a sensitivity analysis.

3 Identification

In this section, we show the identifications for the subpopoulations TBR and THR under non-separable GLMM for both continuous and binary outcomes. Specifically, we have the following model for continuous outcomes

$$\begin{cases} Y_t = \alpha_{t,0} + \alpha_{t,1}^T X + \alpha_{t,2} U + \alpha_{t,3}^T XU + \epsilon_t, \\ \epsilon_t \perp (X, U), \epsilon_t \sim N(0, \sigma_t^2), U \sim N(\mu_U, \sigma_U^2), \alpha_{t,3} \neq 0, \end{cases} \tag{1}$$

where $\alpha_{t,1} = (\alpha_{t,1}^{(1)}, \cdots, \alpha_{t,1}^{(p)})^T, \alpha_{t,3} = (\alpha_{t,3}^{(1)}, \cdots, \alpha_{t,3}^{(p)})^T$ and $t = 0, 1$. Without loss of generality, we take $\alpha_{t,2} > 0$, since otherwise set $U^* = \text{sign}(\alpha_{t,2}) \cdot U$ and $\alpha_{t,2}^* = \text{sign}(\alpha_{t,2}) \cdot \alpha_{t,2}$, where $\text{sign}(k)$ denotes the sign of k. If $\alpha_{t,3} \neq 0$, the model is not separable, i.e., the model cannot be written in the form of $Y_t = l_1(X) + l_2(U)$. This is not a stringent assumption when the observed covariate X is high dimensional since it requires at least one, but not all, interactions between X and U. We will show that, although U is a latent variable, the non-separability assumption can be empirically tested. Under the GLMM (1), we can also empirically test whether Assumption 2 is more reasonable than Assumption 1.

The latent variable U can be interpreted as a subject-specific random effect and the distribution of U is assumed normal. Without loss of generality, we take $(\mu_U, \sigma_U^2) = (0, 1)$ since

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

otherwise U can be standardized. We evaluate the performance of the proposed estimators when the normality of U is violated with a sensitivity analysis in Section 5.2.

We have the following formulas (F1) and (F2) for $\mathrm{TBR}_c(x)$ and $\mathrm{THR}_c(x)$, the proofs of which are given in Appendix A. From (F1) and (F2), once the parameters in model (1) are identified, $\mathrm{TBR}_c(x)$ and $\mathrm{THR}_c(x)$ can be identified. Specifically,

$$\mathrm{TBR}_c(x) = \Phi\Big(\frac{(\alpha_{1,0}-\alpha_{0,0})+(\alpha_{1,1}-\alpha_{0,1})^T x - c}{\sqrt{\big((\alpha_{1,2}-\alpha_{0,2})+(\alpha_{1,3}-\alpha_{0,3})^T x\big)^2+\sigma_0^2+\sigma_1^2}}\Big), \tag{F1}$$

$$\mathrm{THR}_c(x) = \Phi\Big(\frac{(\alpha_{0,0}-\alpha_{1,0})+(\alpha_{0,1}-\alpha_{1,1})^T x - c}{\sqrt{\big((\alpha_{0,2}-\alpha_{1,2})+(\alpha_{0,3}-\alpha_{1,3})^T x\big)^2+\sigma_0^2+\sigma_1^2}}\Big), \tag{F2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable.

When outcomes are binary, we consider the following model

$$\begin{cases} Y_t^* = \alpha_{t,0} + \alpha_{t,1}^T X + \alpha_{t,2} U + \alpha_{t,3}^T XU + \epsilon_t, \\ Y_t = I(Y_t^* > 0), \\ \epsilon_t \perp (X,U), \epsilon_t \sim N(0,\sigma_t^2), U \sim N(\mu_U, \sigma_U^2), (\alpha_{t,0},\alpha_{t,1}) \neq 0, \alpha_{t,3} \neq 0, \end{cases} \tag{2}$$

for $t = 0, 1$, where $\alpha_{t,1} = (\alpha_{t,1}^{(1)}, \cdots, \alpha_{t,1}^{(p)})^T, \alpha_{t,3} = (\alpha_{t,3}^{(1)}, \cdots, \alpha_{t,3}^{(p)})^T$. Again, we assume that U is standard normal. Additionally, without loss of generality, we assume $\sigma_0^2 = \sigma_1^2 = 1$ since otherwise set $\tilde{Y}_t^* = Y_t^*/\sigma_t$, $\tilde{\alpha}_{t,k} = \alpha_{t,k}/\sigma_t$ and $\tilde{\epsilon}_t = \epsilon_t/\sigma_t$ for $t = 0, 1$ and $k = 0, \ldots, 3$. Here Y_t^* is a latent variable and (2) indicates a probit model for the outcome Y_t:

$$P(Y_t = 1 | X, U) = \Phi\Big(\alpha_{t,0} + \alpha_{t,1}^T X + \alpha_{t,2} U + \alpha_{t,3}^T XU\Big).$$

Similar to the continuous case, under (2) we can empirically evaluate whether Assumption 2 is more reasonable than Assumption 1 and whether the GLMM is separable despite U is unobserved.

We have the following formulas (F3) and (F4) for $\mathrm{TBR}(x)$ and $\mathrm{THR}(x)$, the proofs of which are given in Appendix A. From (F3) and (F4), once the parameters in model (2) are identified, $\mathrm{TBR}(x)$ and $\mathrm{THR}(x)$ can also be identified as

$$\mathrm{TBR}(x) = \Phi_b\big(\mu(x;\theta), \Sigma(x;\theta)\big), \tag{F3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

$$\mathrm{THR}(x) = \Phi_h\big(\mu(x;\theta), \Sigma(x;\theta)\big), \tag{F4}$$

where

$$\mu(x;\theta) = \big(\mu_0(x;\theta), \mu_1(x;\theta)\big) = \big(-\alpha_{0,0} - \alpha_{0,1}^T x,\ \ -\alpha_{1,0} - \alpha_{1,1}^T x\big),$$

$$\Sigma(x;\theta) = \begin{pmatrix} 1 + (\alpha_{0,2} + \alpha_{0,3}^T x)^2 & (\alpha_{0,2} + \alpha_{0,3}^T x)(\alpha_{1,2} + \alpha_{1,3}^T x) \\ (\alpha_{0,2} + \alpha_{0,3}^T x)(\alpha_{1,2} + \alpha_{1,3}^T x) & 1 + (\alpha_{1,2} + \alpha_{1,3}^T x)^2 \end{pmatrix},$$

$$\Phi_b\big(\mu, \Sigma\big) = \Phi_2\big((0,\infty), (-\infty, 0); \mu, \Sigma\big),$$

$$\Phi_h\big(\mu, \Sigma\big) = \Phi_2\big((-\infty, 0), (0,\infty); \mu, \Sigma\big),$$

and

$$\Phi_2(A_0, A_1; \mu, \Sigma) = \int\int_{A_0 \times A_1} \frac{1}{2\pi|\Sigma|^{1/2}} \exp\big\{ -\frac{1}{2}(s_0 - \mu_0, s_1 - \mu_1)\Sigma^{-1}(s_0 - \mu_0, s_1 - \mu_1)^T\big\} ds_0 ds_1.$$

Let $\theta = (\theta_0^T, \theta_1^T)^T$ denote the parameters in models (1) and (2), $\theta_t = (\alpha_{t,0}, \alpha_{t,1}^T, \alpha_{t,2}, \alpha_{t,3}^T, \sigma_t^2)^T$ in the continuous model and $\theta_t = (\alpha_{t,0}, \alpha_{t,1}^T, \alpha_{t,2}, \alpha_{t,3}^T)^T$ in the binary model, for $t = 0, 1$. We have the following theorem for the identification of θ and thus the identification of $(\mathrm{TBR}_c(x),\ \mathrm{THR}_c(x))$ for the continuous outcomes and $(\mathrm{TBR}(x),\ \mathrm{THR}(x))$ for the binary outcomes, the proof of which is given in Appendix B.

Theorem 1. *Under Assumption 2,*

(i) If the model (1) holds, the parameters θ can be identified, thus the $\mathrm{TBR}_c(x)$ and $\mathrm{THR}_c(x)$ can also be identified for any constant c.

(ii) If the model (2) holds, the regularity Condition C (given in Appendix B) holds, then the parameters θ can be identified, thus the $\mathrm{TBR}(x)$ and $\mathrm{THR}(x)$ can also be identified.

The non-separable condition plays an important role in the identification of the parameters θ in the presence of U. When marginalizing over U, the interaction term between U and X helps identify the effect of U on Y which would otherwise be absorbed in the intercept. Although the identification of $\alpha_{t,0}$, $\alpha_{t,1}$ and σ_t are straightforward, as shown in the Appendix B, it is $\alpha_{t,2}$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

$\alpha_{t,3}$ that captures the unmeasured heterogeneous treatment effect and allow the identification of TBR and THR. Since U is not measured, the identification of $\alpha_{t,2}$ and $\alpha_{t,3}$ relies on borrowing information from the observed data. Intuitively, this cannot be achieved when $\alpha_{t,3} = 0$ since $\alpha_{t,2}$ will be absorbed in the variance of ϵ_t. In Appendix C, we provide proof of non-identification of TBR and THR for both discrete and continuous outcomes when the interaction between X and U is absent.

Theorem 1 implies the identification of the parameters θ in models (1)–(2). Thus both the non-separability assumption and Assumption 2 can be empirically evaluated using the observed data. We can empirically test for the interaction between X and U, i.e., whether $\alpha_{t,3}$ is significantly different from 0. We can also empirically test the inclusion of U in the models (1)–(2) by checking whether the coefficients $\alpha_{t,2}$ and $\alpha_{t,3}$ are significant. If there is at least one of them significant in the models for both Y_0 and Y_1, then the Assumption 1 is violated and we must include a latent U to make the conditional independence of Y_0 and Y_1 hold.

The covariate X is linear in models (1) and (2). This is not required and is imposed for ease of illustration. When the outcomes are continuous, the general form of the GLMM is

$$\begin{cases} Y_t = g_t(X) + Uh_t(X) + \epsilon_t, \\ \epsilon_t \perp (X, U), \epsilon_t \sim N(0, \sigma_t^2), U \sim N(0,1),\ h_t(0) > 0, \end{cases} \tag{3}$$

for $t = 0, 1$. Similarly, when the outcomes are binary, the general form of the GLMM is

$$\begin{cases} Y_t^* = g_t(X) + Uh_t(X) + \epsilon_t, \\ Y_t = I(Y_t^* > 0), \\ \epsilon_t \perp (X, U), \epsilon_t \sim N(0,1), U \sim N(0,1),\ h_t(0) > 0, \end{cases} \tag{4}$$

for $t = 0, 1$. Models (1) and (2) are special cases of models (3) and (4) with $g_t(X) = \alpha_{t,0} + \alpha_{t,1}^T X$ and $h_t(X) = \alpha_{t,2} + \alpha_{t,3}^T X$. In Appendix B, we give necessary and sufficient conditions to identify $\theta = (g_0(X), h_0(X), \sigma_0^2, g_1(X), h_1(X), \sigma_1^2)$ in model (3) and $\theta = (g_0(X), h_0(X), g_1(X), h_1(X))$ in

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

model (4). Once $(h_0(X), h_1(X))$ is identified, we can test the inclusion of U by testing whether $(h_0(X), h_1(X))$ is significant with the observed data, and test the non-separability by testing whether $(h_0(X), h_1(X)) = (h_0, h_1)$, where h_0 and h_1 are constants.

4 Inference

When the outcomes are continuous, the parameters θ can be estimated by the MLE $\widehat{\theta}$, obtained by maximizing the log-likelihood

$$\ell = \log L(T, X, Y) = P_n\big\{\psi(T, X, Y; \theta)\big\},$$

where $P_n g(X) = \sum_{i=1}^n g(X_i)/n$, and

$$\begin{aligned} & \psi(T, X, Y; \theta) \\ = & \sum_{t=0,1} \frac{1}{2}\bigg[I(T=t)\Big\{ -\log(2\pi) - \log\big((\alpha_{t,2} + \alpha_{t,3}^T X)^2 + \sigma_t^2\big) - \frac{(Y - \alpha_{t,0} - \alpha_{t,1}^T X)^2}{(\alpha_{t,2} + \alpha_{t,3}^T X)^2 + \sigma_t^2}\Big\}\bigg]. \end{aligned}$$

By the theory of M-estimators, we have the asymptotic normality of $\widehat{\theta}$, which can be used to test the significance of the parameters. Additionally, following (F1) and (F2), we can estimate the TBR_c and THR_c by

$$\widehat{\mathrm{TBR}}_c(x) = \Phi\Big(\frac{(\widehat{\alpha}_{1,0} - \widehat{\alpha}_{0,0}) + (\widehat{\alpha}_{1,1} - \widehat{\alpha}_{0,1})^T x - c}{\sqrt{\big((\widehat{\alpha}_{1,2} - \widehat{\alpha}_{0,2}) + (\widehat{\alpha}_{1,3} - \widehat{\alpha}_{0,3})^T x\big)^2 + \widehat{\sigma}_0^2 + \widehat{\sigma}_1^2}}\Big),$$
$$\widehat{\mathrm{THR}}_c(x) = \Phi\Big(\frac{(\widehat{\alpha}_{0,0} - \widehat{\alpha}_{1,0}) + (\widehat{\alpha}_{0,1} - \widehat{\alpha}_{1,1})^T x - c}{\sqrt{\big((\widehat{\alpha}_{0,2} - \widehat{\alpha}_{1,2}) + (\widehat{\alpha}_{0,3} - \widehat{\alpha}_{1,3})^T x\big)^2 + \widehat{\sigma}_0^2 + \widehat{\sigma}_1^2}}\Big),$$

where $\widehat{\alpha}_{t,k}$ and $\widehat{\sigma}_t^2$ are MLEs of the corresponding parameters. The following theorem shows the $\sqrt{n}$ consistency, asymptotic normality and provides the asymptotic variances of the estimators when the outcomes are continuous.

Theorem 2. *If the model (1) holds for continuous outcomes, we have*

$$\sqrt{n}\big(\widehat{\mathrm{TBR}}_c(x) - \mathrm{TBR}_c(x)\big) \xrightarrow{d} N(0, \sigma_{cB}^2(x; \theta)),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

$$\sqrt{n}\big(\widehat{\mathrm{THR}}_c(x) - \mathrm{THR}_c(x)\big) \xrightarrow{d} N(0, \sigma^2_{cH}(x;\theta)),$$

where $\xrightarrow{d}$ *denotes convergence in distribution. The expressions and consistent estimators of* $\sigma^2_{cB}(x;\theta)$ *and* $\sigma^2_{cH}(x;\theta)$ *are given in Appendix D.*

When the outcomes are binary, parameters θ can be estimated by the MLE $\widehat{\theta}$, which is obtained by maximizing the log-likelihood,

$$\ell = \log L(T, X, Y) = P_n\big\{\psi(T, X, Y; \theta)\big\},$$

and

$$\psi(T, X, Y; \theta) \quad = \quad \sum_{t=0,1} \Big[I(T = t)\Big\{ Y \log\big(G(X;\theta_t)\big) + (1 - Y) \log\big(1 - G(X;\theta_t)\big)\Big\}\Big],$$

where

$$G(X;\theta_t) = \Phi\Big(\frac{\alpha_{t,0} + \alpha^T_{t,1} X}{\sqrt{1 + (\alpha_{t,2} + \alpha^T_{t,3} X)^2}}\Big).$$

Similarly, we have the asymptotic normality of $\widehat{\theta}$, which can be used to test the significance of the parameters. Additionally, following (F3) and (F4), we can estimate the TBR and THR by

$$\widehat{\mathrm{TBR}}(x) = \Phi_b\big(\mu(x;\widehat{\theta}), \Sigma(x;\widehat{\theta})\big),$$

$$\widehat{\mathrm{THR}}(x) = \Phi_h\big(\mu(x;\widehat{\theta}), \Sigma(x;\widehat{\theta})\big).$$

The following theorem shows the $\sqrt{n}$ consistency, asymptotic normality and provides the asymptotic variances of the estimators when the outcomes are binary.

Theorem 3. *If the model (2) holds for binary outcomes and the regularity Condition C holds, we have*

$$\sqrt{n}\big(\widehat{\mathrm{TBR}}(x) - \mathrm{TBR}(x)\big) \xrightarrow{d} N(0, \sigma^2_{bB}(x;\theta)),$$

$$\sqrt{n}\big(\widehat{\mathrm{THR}}(x) - \mathrm{THR}(x)\big) \xrightarrow{d} N(0, \sigma^2_{bH}(x;\theta)),$$

where the expressions and consistent estimators of $\sigma^2_{bB}(x;\theta)$ *and* $\sigma^2_{bH}(x;\theta)$ *are given in the Appendix E.*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

5 Simulation

5.1 Finite sample performance

We first assess the finite sample performance of the estimators proposed in Section 4. The simulations were conducted with (a) the continuous outcomes and (b) the binary outcomes. For scenario (a), the simulation study was conducted as follows.

Step 1: A population of size 1000 was created. Variables T, X and U were generated independently. Treatment T was generated from a Bernoulli distribution with $P(T=1)=0.5$, the components of covariates $X=(X_1,X_2,X_3)^T$ were identically and independently generated from a standard normal distribution and latent variable U was also generated from a standard normal distribution. Potential outcomes (Y_0,Y_1) were generated from (1) with parameters set to

$$(\alpha_{0,0},\alpha_{0,1}^{(1)},\alpha_{0,1}^{(2)},\alpha_{0,1}^{(3)},\alpha_{0,2},\alpha_{0,3}^{(1)},\alpha_{0,3}^{(2)},\alpha_{0,3}^{(3)})=(-0.3,1.2,-1.0,-0.8,0.7,-0.5,1.3,0.6),$$

$$(\alpha_{1,0},\alpha_{1,1}^{(1)},\alpha_{1,1}^{(2)},\alpha_{1,1}^{(3)},\alpha_{1,2},\alpha_{1,3}^{(1)},\alpha_{1,3}^{(2)},\alpha_{1,3}^{(3)})=(0.2,-0.8,1.2,1.0,0.8,-0.6,1.0,0.6),$$

$$\sigma_0^2=1.0,\ \sigma_1^2=1.2.$$

Step 2: The parameters θ were estimated using MLE and the estimates of $(\mathrm{TBR}_c(x_{0.25}),\mathrm{THR}_c(x_{0.25}))$ and the variances of the estimators were calculated, where $c=1$, $x_{0.25}=(x_{1,0.25},x_{2,0.25},x_{3,0.25})$ and $x_{i,0.25}$ is the value of the first quartile of the i^{th} covariate distribution.

Step 3: Steps 1 and 2 were repeated for 1000 times to obtain the biases, average estimated standard error (ASE) and the empirical standard error (ESE).

The results where U follows a normal distribution are reported in Table 1. The biases are -0.002 and 0.001 for $\mathrm{TBR}_c(x_{0.25})$ and $\mathrm{THR}_c(x_{0.25})$, respectively, and the ASEs are 0.027 and 0.035

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

respectively (both approximate their ESEs which are 0.028 and 0.036). The coverages of the 95% CI approximate 0.95, indicating good performance of our estimators. We also carried out simulations for the $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ at $x_{0.5}$ and $x_{0.75}$, where $x_{0.5}$ and $x_{0.75}$ are the the median and third quartile of the corresponding covariates distributions. We observed similar results. They are not shown here due to space constraints.

For binary outcomes, the simulation process was similar, except in Step 1, we set the sample size to be 2000 and generated (Y_0, Y_1) from (2) with the same θ excluding (σ_0^2, σ_1^2) and in Step 2, the $(\text{TBR}(x_{0.25}), \text{THR}(x_{0.25}))$ were calculated instead of the $(\text{TBR}_c(x_{0.25}), \text{THR}_c(x_{0.25}))$. The results for the binary outcomes where U was simulated from a normal distribution are shown in Table 2. The biases are -0.001 and 0.001 for $\text{TBR}(x_{0.25})$ and $\text{THR}(x_{0.25})$, respectively, and the ASEs are 0.031 and 0.048, respectively (both approximate their ESEs which are 0.031 and 0.047). The coverages of the 95% CI approximate 0.95, indicating good performance of our estimators.

5.2 Sensitivity analysis with respect to the distribution of U

We assumed that U was normally distributed for the identification of the joint distribution of (Y_0, Y_1). We carried out a sensitivity analysis to evaluate the performance of the estimators for $\text{TBR}(x_{0.25})$, $\text{THR}(x_{0.25})$ with U distributed as t, chi-squared, Poisson and Bernoulli. The estimation was carried out as in Section 5.1 except U was generated from the distributions above. We standardized U to have mean 0 and variance 1 under each distribution.

The results for continuous outcomes are shown in Table 1. When U follows a t(3) distribution, the biases are 0.001 and 0.001 for $\text{TBR}_c(x_{0.25})$ and $\text{THR}_c(x_{0.25})$ respectively, and the ASEs are 0.027 and 0.029, respectively, where the ESEs are 0.028 and 0.029, respectively. The coverages of 95% CI are 0.942 and 0.945, respectively. Similar performance is also observed when U follows distributions such as chi-squared, Poisson and Bernoulli. As the degrees of freedom increase for

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

Table 1: The true value, bias, average estimated standard error (ASE), empirical standard error (ESE), and 95% confidence interval (CI) coverage in scenario (a) with continuous outcomes. Each table cell contains two elements, which corresponds to $\mathrm{TBR}_c(x_{0.25})$ (first row in each cell) and $\mathrm{THR}_c(x_{0.25})$ (second row in each cell) ($c = 1$), respectively.

Distribution of U	true value	bias	ASE	ESE	95% CI coverage
Normal	0.201	-0.002	0.027	0.028	0.947
	0.585	0.001	0.035	0.036	0.945
t(3)	0.288	0.001	0.027	0.028	0.942
	0.470	0.001	0.029	0.029	0.945
t(10)	0.214	0.002	0.027	0.029	0.938
	0.564	-0.001	0.034	0.035	0.946
$\chi^2(3)$	0.182	0.001	0.029	0.029	0.947
	0.612	0.001	0.037	0.036	0.954
$\chi^2(10)$	0.182	-0.001	0.028	0.029	0.943
	0.612	0.002	0.037	0.037	0.952
P(3)	0.234	-0.001	0.028	0.028	0.953
	0.540	-0.002	0.033	0.033	0.955
P(10)	0.215	-0.002	0.028	0.028	0.943
	0.564	0.002	0.034	0.034	0.947
B(0.5)	0.112	-0.001	0.025	0.025	0.954
	0.728	0.001	0.04	0.039	0.955

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

Table 2: The true value, bias, average estimated standard error (ASE), empirical standard error (ESE), and 95% confidence interval (CI) coverage in scenario (b) with binary outcomes. Each table cell contains two elements, which corresponds to TBR($x_{0.25}$) (first row in each cell) and THR($x_{0.25}$) (second row in each cell), respectively.

Distribution	true value	bias	ASE	ESE	95% coverage
Normal	0.109	-0.001	0.031	0.031	0.937
	0.420	0.001	0.048	0.047	0.948
t(3)	0.174	-0.004	0.025	0.023	0.956
	0.315	0.003	0.040	0.040	0.935
t(10)	0.123	-0.002	0.030	0.028	0.963
	0.400	0.001	0.048	0.046	0.951
$\chi^2(3)$	0.094	0.003	0.030	0.027	0.958
	0.443	0.012	0.053	0.050	0.937
$\chi^2(10)$	0.095	0.003	0.031	0.030	0.947
	0.442	0.007	0.051	0.050	0.945
P(3)	0.136	-0.001	0.029	0.028	0.950
	0.379	0.001	0.050	0.047	0.955
P(10)	0.121	0.001	0.031	0.030	0.950
	0.403	0.001	0.050	0.049	0.942
B(0.5)	0.039	0.025	0.037	0.040	0.836
	0.513	-0.026	0.052	0.050	0.952

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

distributions such as chi-squared and Poisson, the standardized U can be approximated by normal and the good performances of estimators is expected. When the degrees of freedom are small, the performance of estimators are robust for symmetric distributions (e.g., t-distribution) and skewed distributions (e.g., chi-squared distribution). The estimators are robust even for the discrete distributions (Poisson, Bernoulli).

The results of the binary case are in Table 2. When U follows a t(3) distribution, the biases are -0.004 and 0.003 for $\text{TBR}(x_{0.25}), \text{THR}(x_{0.25})$, respectively, and the ASE are 0.025 and 0.040, respectively (the ESEs which are 0.023 and 0.040, respectively). The coverages are 0.956 and 0.935, respectively, which approximate 0.95. From the table we infer that when the outcomes are binary, the estimators are robust to the different distributions of the unmeasured variable U, including symmetric, non-symmetric, and discrete distributions.

We carried out sensitivity analysis for different distributions of U for the conditional heterogenous treatment effects at $x_{0.5}$ and $x_{0.75}$, as well as the marginal heterogeneous treatment effects under scenario (a) and (b). We observed similar results, which were not shown here.

6 Statistical Analysis

6.1 MIND Study

The Memory in Diabetes (MIND) study was the first randomized trial in older persons with type 2 diabetes to test the effect of intensive compared to standard glycaemic therapeutic strategies on multiple cognitive domains and on structural changes in the brain (**?**). People with type 2 diabetes are at risk for cognitive impairment and brain atrophy. The study participants were randomized to an intensive glycaemic therapeutic strategy targeting HbA_{1c} to <6%, or a standard strategy targeting HbA_{1c} to 7%–7.9%. Of the 614 participants with a baseline MRI, 230 intensive and 273

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

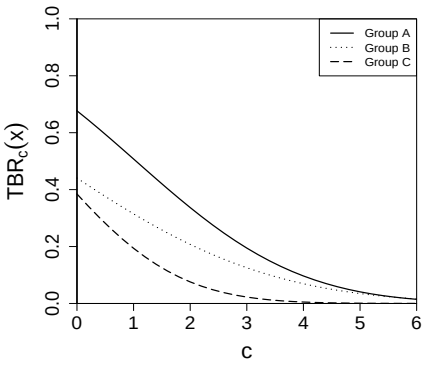
standard therapy participants were included in the analysis. Our primary outcome is the abnormal white matter (AWM) tissue volume at 40 months, which reflects diffuse and focal ischemic, demyelinating, and inflammatory processes leading to small vessel disease, and is associated with diabetes and impaired cognition (**??**).

We calculated the ATE among different subgroups. We adjusted for gender (male or female), race (white or not), history of CVD (yes or no), age (< 60, 60–69, 70+ yrs), and the number of correctly completed cells on the 40-month Digit Symbol Substitution Test (DSST) (< 47, 47–59, 60+), as suggested in **?**. For illustration, we focus on three subgroups: group A is nonwhite females under 60 years old without CVD history and DSST < 47; group B is white male under 60 years with CVD history and DSST > 60, and group C is nonwhite males over 70 years old without CVD history and DSST < 47. Amongst the subgroups defined by the available covariates, the only subgroup with a significant ATE is group A: $\widehat{ATE}(x_A) = 1.043$, 95% CI $[0.102, 1.984]$, p-value $= 0.030$. Groups B and C have negative ATEs with similar p-values: $\widehat{ATE}(x_B) = -0.449$, 95% CI $[-1.605, 0.707]$, p-value $= 0.447$, and $\widehat{ATE}(x_C) = -0.514$, 95% CI $[-1.813, 0.785]$, p-value $= 0.439$. However, with the average mean treatment effect for each subgroup, additional information is needed to further describe the treatment effect. For example, what proportion of individuals in group A that benefit from the treatment, and what is the distribution of such benefit. Although groups B and C have negative ATE, are there any individuals benefit from the treatment in these groups, and how much harm does the treatment cause in these groups.

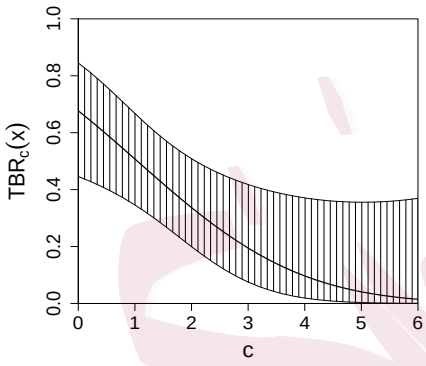
To further estimate the $\text{TBR}_c(x)$ and $\text{THR}_c(x)$, we adjusted in model (2) for the same set of covariates when calculating $\text{ATE}(x)$. The results of the regression suggest some significant interactions, as shown in Table 2 of the Appendix H (e.g., the coefficient of Ugender is estimated to be $\widehat{\alpha}_{0,1}^{(1)} = -1.916$, 95% CI: $[-2.856, -0.975]$, p-value < 0.001, $\widehat{\alpha}_{1,1}^{(1)} = -1.742$, 95% CI: $[-2.553, 0.931]$, p-value < 0.001). This justifies the inclusion of U in the model (2) and the non-separable assump-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

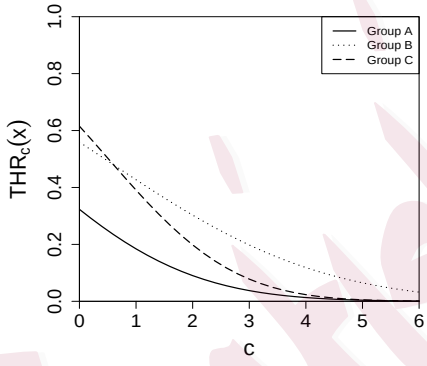
Figure 1: Estimated curves of $\mathrm{TBR}_c(x)$ and $\mathrm{THR}_c(x)$ of MIND study



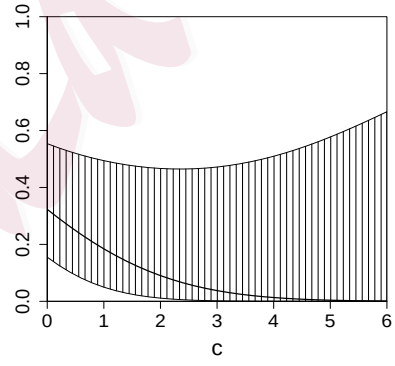
(a) Estimated curve of $\mathrm{TBR}_c(x)$ in different subgroups



(b) Estimated curve of $\mathrm{TBR}_c(x)$ and its confidence band in group A



(c) Estimated curve of $\mathrm{THR}_c(x)$ in different subgroups



(d) Estimated curve of $\mathrm{THR}_c(x)$ and its confidence band in group A

Note: Group A is nonwhite females under 60 years old without CVD history, DSST < 47; group B is white male under 60 years with CVD history, DSST > 60; group C is nonwhite males over 70 years old without CVD history, DSST < 47.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

tion holds. The estimated curves of $\text{TBR}_c(x)$ and $\text{THR}_c(x)$ with different threshold values c are given in Figure 1 (a) and (c). Each line stands for a subgroup with specific covariate values. A comparison of the TBR and THR curves for group A with groups B and C reveals individuals in group A benefit the most from the treatment as compared with groups B and C since the $\text{TBR}_c(x)$ is larger and $\text{THR}_c(x)$ is smaller for group A at all levels of c. This information goes beyond the mean treatment effect being larger in group A as compared with group B and C.

Due to the space constraint, we show only the confidence band of the TBR_c and THR_c for group A in Figure 1 (b) and (d). Although the $\text{ATE}(x_A)$ is significantly different from 0, the individual treatment effect is quite heterogenous: about two thirds of individuals benefit from it ($\widehat{\text{TBR}}_{c=0}(x_A) = 0.677$, 95% CI $[0.467, 0.886]$, p-value < 0.001) and approximately a third of individuals in this subgroup are harmed by it ($\widehat{\text{THR}}_{c=0}(x_A) = 0.323$, 95% CI $[0.113, 0.533]$, p-value < 0.001). Additionally, there are about 20% of individuals having a relative large positive treatment effect, e.g., $\widehat{\text{TBR}}_{c=3}(x_A) = 0.195$, 95% CI $[0.024, 0.365]$, p-value $= 0.025$. Thus, the covariates in the models do not fully describe the individual characteristics that would benefit from the treatment and it would harm a third of individuals if the treatment is advocated uniformly in group A.

Although groups B and C have similar negative ATEs with comparable standard errors, the treatment has quite different impact on the two groups. For group B, there are 12.5% of individuals having a relatively large positive treatment effect ($\widehat{\text{TBR}}_{c=3}(x_B) = 0.125$, 95% CI $[0.017, 0.233]$, p-value < 0.001) while that is only 2.2% for group C ($\widehat{\text{TBR}}_{c=3}(x_C) = 0.022$, 95% CI $[-0.021, 0.065]$, p-value $= 0.317$). These results suggest group C may have a more homogenous treatment effect as compared with group B. Since some individuals in group B have a relatively large treatment effect, it may be worthwhile to incoporate more information and subject matter knowledge to identify these individuals. These information will all be missed if we only investigate

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

the subgroup ATE.

6.2 ACCORD Eye Study

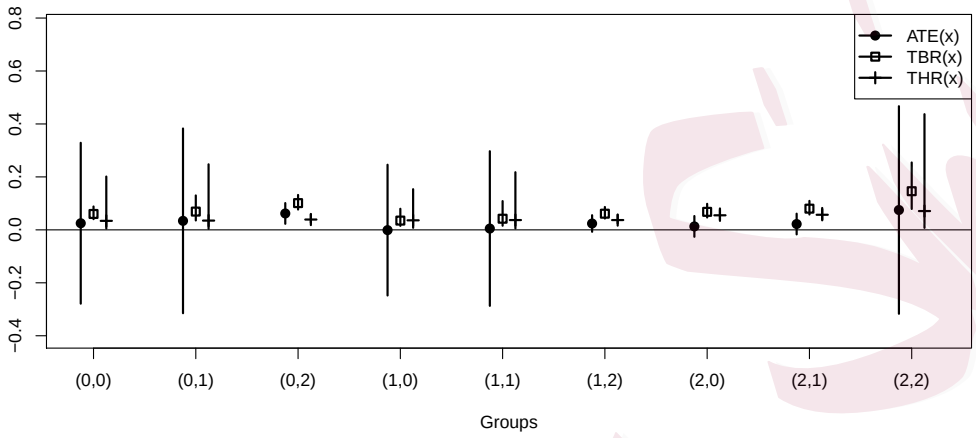
Diabetic retinopathy (DR) is one of the most common causes of vision loss among people with diabetes, and the leading cause of vision impairment and blindness among adults younger than 65 in the United States (Chew, Ambrosius, Howard, Greven, Johnson, Danis, Davis, Genuth, and Domanski (2007)). It has been found that poor glycemic control is one of the most important risk factors associated with the development of DR. The Action to Control Cardiovascular Risk in Diabetes Eye (ACCORD Eye) study aimed at determining whether the intensive glycemia control could reduce the risk of development or progression of diabetic retinopathy, as compared with the standard treatments. The study included 2856 eligible participants randomly assigned to either intensive or standard treatment for glycemia. The primary outcome for this study was the composite end point of either progression of diabetic retinopathy by at least three steps on the Early Treatment Diabetic Retinopathy Study (ETDRS) Severity Scale, or development of proliferative diabetic retinopathy necessitating photocoagulation therapy or vitrectomy in 4 years ($Y = 0$ if the progression of diabetic retinopathy was seen, and $Y = 1$ otherwise) (Group and Group (2010)).

At the end of four years of follow-up, progression of diabetic retinopathy was seen in 7.3% of participants (104 out of 1429) in the intensive glycemic control group, as compared with 10.4% of participants (149 out of 1427) in the standard glycemic therapy group, suggesting a significant effect of the intensive glycemic control for the entire study population ($\widehat{\text{ATE}} = 0.032$, 95% CI $[0.011, 0.052]$, p-value $= 0.003$).

To investigate how the ATE changes across different subgroups, we carried out a subgroup analysis. We adjusted for the other treatment that individuals used and the qualification to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

Figure 2: Estimates and confidence intervals of ATE(x), TBR(x) and THR(x) for different groups $x = (x_1, x_2)$ in ACCORD Eye study



participate in the Lipid and blood pressure (BP) trials, that is, we define subgroups by X_1 (0 if in Lipid placebo arm, 1 if in Lipid Fibrate arm, and 2 if not eligible for the Lipid trial) and X_2 (0 if in standard BP arm, 1 if in intensive BP arm, and 2 if not eligible for the BP trial). Covariates (X_1, X_2) were considered by Group and Group (2010). The estimates and CIs of subgroup ATE are presented in Figure 2. Although the population ATE is significant, as shown above, the only subgroup with a significant ATE is the group $(0, 2)$ ($\widehat{\text{ATE}}(0, 2) = 0.062$, 95% CI $[0.023, 0.101]$, p-value $= 0.002$). The subgroup ATE suggests only using the intensive glycemic control in group $(0, 2)$, that is in individuals not eligible for the BP trial but in the Lipid placebo arm.

To estimate TBR(x) and THR(x), we adjusted in model (2) for (X_1, X_2) as nominal variables. The results of the regression suggest some significant interactions, including $UX_1^{(1)}$ ($\widehat{\alpha}_{0,3}^{(1)} = 0.556$, 95% CI: $[0.238, 0.873]$, p-value $= 0.021$), $UX_1^{(2)}$ ($\widehat{\alpha}_{1,3}^{(2)} = 0.531$, 95% CI: $[0.210, 0.852]$, p-value $= 0.027$), where $X_1^{(1)}$ and $X_1^{(2)}$ are indicators for eligibility to participate in the Lipid trial and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

the use of Lipid Fibrate treatment. This justifies the inclusion of U in the model (2) and the non-separable assumption holds.

The estimates and confidence intervals for TBR(x) and THR(x) in different subgroups are shown in Figure 2. The estimates of both are relatively small across all groups. Specifically, for the individuals not eligible for the BP trial but in the Lipid placebo arm (group $(0, 2)$), $\widehat{\text{TBR}}(0, 2) = 0.101$, 95% CI $[0.074, 0.128]$, p-value < 0.001 and $\widehat{\text{THR}}(0, 2) = 0.039$, 95% CI $[0.023, 0.055]$, p-value < 0.001. Although the subgroup ATE suggests the use of the intensive glycemic control for group $(0, 2)$, the heterogeneity analysis reveals that over 85% of individuals in this subgroup neither benefit nor are harmed. This trend holds for most groups defined by the available covariates (X_1, X_2): over 80% of individuals have no effect at all across all subgroups. On the other hand, although having a non-significant subgroup ATE, group $(2, 2)$ has about 15% individuals benefit from the intensive glycemic control ($\widehat{\text{TBR}}(2, 2) = 0.146$, 95% CI $[0.060, 0.232]$, p-value < 0.001), which is the highest proportion of individuals having benefits among all subgroups. Additional information is needed to further identify these subjects.

7 Discussion

In this article, we assessed the treatment effect heterogeneity by evaluating the TBR and THR. We relaxed the conditional independence Assumption 1 by allowing the presence of an unmeasured latent variable. Under non-separable (generalized) linear mixed models, the existence of the latent variable can be tested, and we provided identification and estimation methods.

We imposed a normality assumption on the latent variable U. Normality of U is not necessary for identification, but when the distribution of U is not normal, the distribution of $Y - g(X)$, conditional on X, may not have a distribution in closed form and the identification condition may thus be complicated. We carried out a sensitivity analysis to evaluate the performance of estimators under different underlying distributions of U. We leave the generalization of identification and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

estimation of treatment effect heterogeneity under different distributions of U as future research topics.

The identification and inference method we developed in this paper rely on the parametric assumptions. They have efficiency gains and result in better convergence as compared with a semiparametric approach, but are not as robust. We leave the development of a semiparametric approach for future research.

8 Supplementary Material

In Appendix A, we provide the proof of formulas (F1), (F2), (F3), and (F4) and derive the relationship of ATE(x), TBR$_c(x)$ and THR$_c(x)$. In Appendix B, we provide the sufficient and necessary identification conditions for $(g_t(X); h_t(X))$ in the models (3) and (4), and prove Theorem 1. In Appendix C, we prove that the treatment effect heterogeneity cannot be identified in separable models. In Appendices D and E, we prove Theorems 2 and 3. In Appendix F, we provide the estimation and asymptotic properties for the models (3) and (4). In Appendix G, we provide identification conditions when U depends on X. In Appendix H, there are additional tables from simulation study and statistical analysis.

Acknowledgements

The content is solely the responsibility of the authors. The authors thank Professor Lan Wang and Professor Xiao-Hua Zhou for fruitful discussions, and Professor Andrew J. Vickers for his generosity in making data available. The authors thank an associate editor and the reviewer for insightful suggestions which have significantly improved the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

References

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(5):662–673.

Bradley, H., Mayosi, B., Maroney, R., Mbewu, A., Opie, L., and Volmink, J. (2007). Beta-blockers for hypertension. *Cochrane Database of Systematic Reviews*, 24:CD002003.

Brand, J. E. and Thomas, J. S. (2013). Causal effect heterogeneity. In *Handbook of Causal Analysis for Social Research*, pages 189–213. Springer.

Chew, E. Y., Ambrosius, W. T., Howard, L. T., Greven, C. M., Johnson, S., Danis, R. P., Davis, M. D., Genuth, S., Domanski, M., Group, A. S., et al. (2007). Rationale, design, and methods of the action to control cardiovascular risk in diabetes eye study (accord-eye). *The American Journal of Cardiology*, 99(12):S103–S111.

Debette, S. and Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ*, 341:c3666.

Gadbury, G. L. and Iyer, H. K. (2000). Unit–treatment interaction and its practical consequences. *Biometrics*, 56(3):882–885.

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(2):163–174.

Gadbury, G. L., Iyer, H. K., and Allison, D. B. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics*, 11(4):313–333.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Launer, L. J., Miller, M. E., Williamson, J. D., Lazar, R. M., Gerstein, H. C., Murray, A. M., Sullivan, M., Horowitz, K. R., Ding, J., Marcovina, S., et al. (2011). Effects of randomization to intensive glucose lowering on brain structure and function in type 2 diabetes accord memory in diabetes study. *Lancet Neurology*, 10(11):969.

Poulson, R. S., Gadbury, G. L., and Allison, D. B. (2012). Treatment heterogeneity and individual qualitative interaction. *The American Statistician*, 66(1):16–24.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Shen, C., Jeong, J., Li, X., Chen, P. S., and Buxton, A. (2013). Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics*, 69(3):724–731.

The ACCORD Study Group and ACCORD Eye Study Group (2010). Effects of medical therapies on retinopathy progression in type 2 diabetes. *New England Journal of Medicine*, 363:233–244.

van Harten, B., de Leeuw, F., Weinstein, H., Scheltens, P., and Biessels, G. (2006). Brain imaging in patients with diabetes. *Diabetes care*, 29:2539–2548.

Yin, Y. and Zhou, X. H. (2016). Using secondary outcome to sharpen inference in characterizing heterogeneity. *Submitted for pulication*.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0150

Yin, Y., Zhou, X. H., Geng, Z., and Lu, F. (2016). Assessing the heterogeneity of treatment effects by identifying the treatment benefit and treatment harm rate. *Submitted for pulication*.

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: Series C (Applied Statistics)*, 62(5):687–704.