BIOSTATISTICS & EPIDEMIOLOGY
https://doi.org/10.1080/24709360.2021.2024972

Taylor & Francis
Taylor & Francis Group

# Nonparametric inference of complier quantile treatment effects in randomized trials with imperfect compliance

Lu Mao

Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA

**ABSTRACT**

To analyze randomized trials with imperfect compliance, a standard approach is to estimate the local average treatment effect in the subpopulation of compliers using randomization status as an instrumental variable. Though quantile analysis has been popular in general, the local (or complier) quantile treatment effect (cQTE) as a causal estimand has received insufficient attention. In this paper, we map out the details for the estimation, inference, and sensitivity analysis of the cQTE in a completely nonparametric setting. We propose to estimate the cQTE using nonparametric plug-in estimators of the cumulative distribution functions for the potential outcomes of the compliers. The cQTE estimator is shown to be asymptotically normal, with asymptotic variance estimated through kernel-smoothed density estimators. The procedure is easily extended to adjust for discrete covariates for gains in statistical efficiency. Moreover, by exploiting the stochastic monotonicity of the quantile functional, we develop sensitivity bounds for the cQTE when key assumptions such as exclusion restriction and instrument monotonicity are violated. Extensive simulations show that the proposed methods provide valid inference of the target local estimand and outperform standard intent-to-treat tests, especially under low compliance rates and/or heterogeneous treatment effects. A recent study on a government-funded health insurance program in India is analyzed as an illustration.

**ARTICLE HISTORY**
Received 1 January 2021
Accepted 15 December 2021

**KEYWORDS**
Causal inference; instrumental variable; kernel density estimator; local treatment effects; sensitivity analysis

## 1. Introduction

As a summary measure, the quantiles are commonly favored over the mean when the outcome distribution is skewed or when treatment effects are heterogeneous [1,2]. Indeed, in these situations, changes in the quantiles can often reveal patterns not seen in the simple average. Given such merits, quantile analysis has been studied thoroughly as a methodological topic [3,4] and applied extensively in fields as diverse as econometrics [5], ecology [6], and medicine [7].

The usual quantile analysis, however, does not address the potential bias due to endogenous treatment or exposure whose relationship with the outcome is confounded by

**CONTACT** Lu Mao lmao@biostat.wisc.edu Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53726, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/24709360.2021.2024972

unknown factors. A common scenario of endogenous treatment comes from randomized trials with imperfect compliance, where some participants, ignoring the randomized assignment, self-select into the treatment of their own choosing. For example, in a recent study to evaluate the effect of a health insurance program funded by the Indian federal government [8], about 25% of the participants who were randomized to the treatment group did not enroll in the program, while about 30% of those randomized to the control enrolled through other means. In order to circumvent the selection bias in the nonrandom treatment, investigators often follow the intent-to-treat (ITT) principle, that is, to analyze the outcomes according to the subject's randomization status rather than the treatment received. The ITT analysis leads to a valid test on the 'sharp' null hypothesis of no treatment effect at the individual level [9], and meanwhile produces an estimate for causal effect of randomization (which is sometimes policy relevant). On the other hand, it does not quantify the causal effect of the treatment *per se*. In fact, without unrealistically strong assumptions, the global treatment effect on the whole population is unidentifiable due to unmeasured confounding [10]. Nevertheless, with the exogenous randomization status as an instrumental variable (IV), local treatment effects on the sub-population of compliers, i.e. those who always follow the assignment, can often be teased out bias-free. For example, under assumptions such as exclusion restriction (that randomization has no direct effect on the outcome) and instrument monotonicity (that randomization influences treatment choice monotonically), Angrist, Imbens, and Rubin [11] derived a nonparametric Wald (IV) estimator for the local (or complier) average treatment effect (ATE) [12–14].

Following this seminal work, a number of authors extended the IV methodology to quantile analysis under endogenous treatment. Their work is mostly set in a general framework where the IV itself is non-randomized (albeit with known confounders). The non-randomized IV necessitates additional assumptions on its conditional probabilities, which often require complex numerical procedures to estimate. For example, Abadie et al. [15,16] considered regression models for the complier quantile treatment effect (cQTE) against an endogenous treatment and other covariates using inverse propensity score weighting for the IV. Due to the noncollapsibility of the quantiles, the cQTE conditional on covariates has no simple correspondence with the (more interpretable) unconditional cQTE, i.e. the quantile difference between the *marginal* distributions of the complier outcomes. Frölich and Melly [17] extended their work to covariate-adjusted unconditional cQTE. The authors proposed kernel-based nonparametric local linear regression to estimate the IV scores and used the estimated scores in a weighted minimization scheme to obtain the cQTE [18]. For randomized trials, although much simpler solutions may be obtained by adapting Abadie et al. (2002; 2003) and Frölich and Melly (2013) under randomized instrument (e.g. via substitution of treatment probability for the propensity score) [19,20], applied researchers, especially medical scientists and practitioners, generally lack the statistical background to implement the simplification and, as a result, hesitate to use these methods in practice. It is thus helpful for statisticians to make the methodology more transparent, in an effort to increase its likelihood of adoption in actual trials.

To take up this task, we examine closely the estimation, inference, and sensitivity analysis of the cQTE in the special setting of randomized trials with noncompliance. In Section 2, we define cQTE estimand, construct an empirical estimator under standard assumptions, and derive its asymptotic variance in analytic form. Exploiting the stochastic monotonicity of the quantile functional, we develop simple and easy-to-compute sensitivity bounds for the estimand when exclusion restriction and instrument monotonicity are violated [21–27]. Simulation studies are conducted in Section 3 to evaluate the performance of the inference procedures and sensitivity bounds and to compare the cQTE with standard ITT analyses in testing the treatment effect. In Section 4, the aforementioned Indian health insurance program study is analyzed using the proposed methodology as an illustration. We conclude the paper with some discussion on future research directions in Section 5.

## 2. Theory and methods

### *2.1. Data and estimand*

We first briefly review the set-up of randomized trials with noncompliance and the definition of the complier quantile treatment effect (cQTE). For a detailed exposition, see Melly and Wüthrich [19, Section 1.2.1].

Let $Z = 1, 0$ denote the randomization status with 1 indicating the treatment and 0 the control. To allow for discrepancy between treatment assignment and receipt, let $A(z) = 1, 0$ denote the potential treatment received had the subject been randomized to group $z$ (z=1, 0) [28]. Under this notation, we can divide the target population into four compliance classes, or principal strata [11,29]. These are always-takers: $A(1) = A(0) = 1$; compliers: $A(1) = 1, A(0) = 0$; never-takers: $A(1) = A(0) = 0$; and defiers: $A(1) = 0, A(0) = 1$. Use $Y(a)$ to denote the potential outcome under treatment $a$ ($a = 1, 0$). Under the Stable Unit Treatment Value Assumption (SUTVA), the actual treatment received is $A = ZA(1) + (1 - Z)A(0)$. Likewise, the observed outcome under received treatment $A$ is $Y = AY(1) + (1 - A)Y(0)$. The observed data thus consist of the triple $(Z, A, Y)$.

We make the following standard assumptions regarding the data-generating mechanism [11].

**(A1)** (Exclusion restriction) If $Y(z, a)$ denotes the potential outcome under randomization status $z$ and treatment $a$, then $Y(z, a) = Y(a)$ with probability 1.
**(A2)** (Randomization) $\{Y(1), Y(0), A(1), A(0)\} \perp\!\!\!\perp Z$.
**(A3)** (Relevance) $\mathbb{P}\{A(1) = 1\} \neq \mathbb{P}\{A(0) = 1\}$.
**(A4)** (Monotonicity) $A(1) \geq A(0)$ with probability one.

In particular, the exclusion restriction assumption implies that randomization affects the outcome only through the treatment received (i.e. no direct effect of randomization). The monotonicity assumption denies the existence of defiers in the population. These assumptions are reasonable in double-blind, placebo-controlled clinical trials but need not always be so in sociological experiments where blinding is impossible (e.g. when subjects are assigned to participate in job training programs [15]). In the latter cases, sensitivity analysis is generally needed to gauge the impact of possible violations against these assumptions (see Section 2.4).

Without further assumptions on the confounding mechanisms between $A$ and $Y$, it is clear that the marginal distributions of the $Y(a)$ cannot be nonparametrically identified. For example, we cannot hope to identify the distribution of $Y(1)$ among the never-takers because we never observe $Y(1)$ in that strata. This means that global treatment effects, such as the difference between the averages or quantiles of $Y(1)$ and $Y(0)$, cannot be estimated from empirical data. There are two general approaches to overcoming this nonidentifiability. The first one is to follow the intent-to-treat (ITT) principle. Let $\omega_z(y) = \mathbb{P}(Y \leq y \mid Z = z)$ denote the cumulative distribution function (cdf) of the observed outcome $Y$ in randomized group $z$ $(z = 1, 0)$. Then, the $\tau$th $(0 < \tau < 1)$ ITT quantile treatment effect is

$$Q_{\mathrm{ITT}}(\tau) = \omega_1^{-1}(\tau) - \omega_0^{-1}(\tau).$$

Here and in the sequel, we adopt the following definition for the 'inverse' function:

$$\nu^{-1}(\tau) = \inf\{y \in \mathbb{R} : \nu(y) \geq \tau\}, \tag{1}$$

which applies to discrete as well as continuous cdfs.

The quantity $Q_{\mathrm{ITT}}(\tau)$ measures the $\tau$-quantile change due to the randomization, not one due to the treatment. Though the latter is not fully identifiable, a local version of it is. Use $\nu_{ac}(y) = \mathbb{P}\{Y(a) \leq y \mid A(1) > A(0)\}$ to denote the potential outcome distributions among the compliers. Then, the complier (or local) $\tau$-quantile treatment effect (cQTE) is

$$Q_{\mathrm{c}}(\tau) = \nu_{1\mathrm{c}}^{-1}(\tau) - \nu_{0\mathrm{c}}^{-1}(\tau), \tag{2}$$

which can be interpreted as the $\tau$-quantile change in the compliers due to the treatment.

By Imbens and Rubin [30], the cdfs $\nu_{ac}$ $(a = 1, 0)$ are indeed identifiable from the observed data. In fact, under assumptions (A1)–(A4), the $\nu_{ac}$ can be represented in the neat form [31]

$$\nu_{ac}(y) = \frac{\mathbb{P}\left(Y \leq y, A = a \mid Z = 1\right) - \mathbb{P}\left(Y \leq y, A = a \mid Z = 0\right)}{\mathbb{P}\left(A = a \mid Z = 1\right) - \mathbb{P}\left(A = a \mid Z = 0\right)}. \tag{3}$$

Consequently we can identify $Q_{\mathrm{c}}(\tau)$ as a direct functional of the $\nu_{ac}(\cdot)$.

## 2.2. Estimation and inference

Given a random $n$-sample of observed data $(Z, A, Y)$, namely,

$$(Z_i, A_i, Y_i) \quad i = 1, \ldots, n, \tag{4}$$

we can construct an estimator $\widehat{\nu}_{ac}(\cdot)$ for $\nu_{ac}(\cdot)$ by replacing the conditional probabilities on the right-hand side of (3) with their empirical analogs. It can be shown that the resulting estimator is equivalent to Frölich and Melly's (2013) weighting estimator when the propensity score is replaced by the corresponding empirical proportions (see supplemental material for details). Plugging in the $\widehat{\nu}_{ac}(\cdot)$ on the right-hand side of (2), we obtain the estimator

$$\widehat{Q}_{\mathrm{c}}(\tau) = \widehat{\nu}_{1\mathrm{c}}^{-1}(\tau) - \widehat{\nu}_{0\mathrm{c}}^{-1}(\tau).$$

We show that $\widehat{Q}_{\mathrm{c}}(\tau)$ is asymptotically normal and derive its asymptotic variance. To do so, we first linearize the $\widehat{\nu}_{1\mathrm{c}}(\cdot)$ by

$$n^{1/2}\{\widehat{\nu}_{ac}(\cdot) - \nu_{ac}(\cdot)\} = n^{-1/2}\sum_{i=1}^{n} \psi_a(Z_i, A_i, Y_i)(\cdot) + o_p(1) \tag{5}$$

for some mean-zero influence function $\psi_a(Z, A, Y)$ (the details are relegated to the online supplemental material). Next, applying the delta method to the 'inverse' functional defined in (1), we obtain that

$$n^{1/2}\{\widehat{\nu}_{ac}^{-1}(\tau) - \nu_{ac}^{-1}(\tau)\} = n^{-1/2}\sum_{i=1}^{n}\eta_a(Z_i, A_i, Y_i)\{\nu_{ac}^{-1}(\tau)\} + o_p(1),$$

where $\eta_a(Z, A, Y)(y) = -\psi_a(Z, A, Y)(y)/\dot{\nu}_{ac}(y)$ and $\dot{\nu}_{ac}(y) = \mathrm{d}\nu_{ac}(y)/\mathrm{d}y$. We then immediately have that

$$n^{1/2}\{\widehat{Q}_c(\tau) - Q_c(\tau)\} \rightarrow_d N\left\{0, \sigma^2(\tau)\right\}, \tag{6}$$

where $\sigma^2(\tau) = E[\eta_1(Z, A, Y)(y) - \eta_0(Z, A, Y)(y)]^2$. Given estimators $\widehat{\dot{\nu}}_{ac}$ and $\widehat{\psi}_a$ for the functions $\dot{\nu}_{ac}$ and $\psi_a$, respectively, we can estimate $\sigma^2(\tau)$ by the moment estimator $\widehat{\sigma}^2(\tau) = n^{-1}\sum_{i=1}^{n}[\widehat{\eta}_1(Z_i, A_i, Y_i)\{\widehat{\nu}_{1c}^{-1}(\tau)\} - \widehat{\eta}_0(Z_i, A_i, Y_i)\{\widehat{\nu}_{0c}^{-1}(\tau)\}]^2$, where $\widehat{\eta}_a(Z, A, Y)(y) = -\widehat{\psi}_a(Z, A, Y)(y)/\widehat{\dot{\nu}}_{ac}(y)$ $(a = 1, 0)$.

The estimator $\widehat{\psi}_a$ can be easily constructed given the form of $\psi_a$ (see the online supplemental material for details). For the density function $\dot{\nu}_{ac}(\cdot)$, direct estimation is impossible since the empirical estimator $\widehat{\nu}_{ac}(y)$ is a step function and hence lacks a well-defined derivative. To avoid this problem, we take a kernel-smoothed approach to estimating $\dot{\nu}_{ac}(y)$ using $\widehat{\nu}_{ac}(y)$. Suppose that $K_h(\cdot)$ is a kernel function satisfying $\int K_h(y)\mathrm{d}y = 1$ with smoothness controlled by a 'bandwidth' parameter $h$. Typically, $K_h(\cdot)$ is the density function for a location-scale family of continuous distributions with dispersion parameter $h$. A common example is the Gaussian kernel $K_h(\cdot - \theta) = h^{-1}\phi\{(\cdot - \theta)/h\}$, where $\phi$ is the density function for the standard normal distribution. Then, a kernel-smoothed estimator for $\dot{\nu}_{ac}(y)$ is

$$\widehat{\dot{\nu}}_{ac}(y) = \int K_{h_{a,n}}\left(y - x\right)\mathrm{d}\widehat{\nu}_{ac}(x), \tag{7}$$

where $h_{a,n}$ is a properly chosen bandwidth parameter satisfying $h_{a,n} \downarrow 0$ as $n \rightarrow \infty$ for each $a = 1$, 0. It is well known that, under suitable regularity conditions, the optimal bandwidth $h_{a,n}$ is of the order $O(n^{-1/5})$. Per the rule of thumb by Silverman [32], we use $h_{a,n} = 1.06\widehat{\sigma}_a n^{-1/5}$, where $\widehat{\sigma}_a$ is the standard deviation of the $Y_i$ with $Z_i = A_i = a$.

**Remark 2.1:** It is clear that $\widehat{\nu}_{ac}(\cdot)$ as an empirical analog of the right-hand side of (3) need not be always nondecreasing in $y$. As a result, the corresponding kernel density estimates can be negative [30]. Obviously, negative density estimates are most likely around points on which $\widehat{\nu}_{ac}(\cdot)$ has a negative jump size. In our application, however, negativity is less of a concern because the estimated density function is evaluated only at $\widehat{\nu}_{ac}^{-1}(\tau)$, which is always associated with a positive jump size (see definition of the inverse function in (1)). Indeed, in all simulations described in Section 3, we have not encountered a single case with negative density estimates.

The nonparametric cQTE estimator $\widehat{Q}_c(\tau)$ can also be used in hypothesis testing. Specifically, an asymptotic level-$\alpha$ $(0 < \alpha < 1)$ test rejects the null hypothesis of no treatment effect if $n^{1/2}|\widehat{Q}_c(\tau)| > \widehat{\sigma}(\tau)z_{1-\alpha/2}$, where $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ and $\Phi(\cdot)$ standard normal cdf. This IV test can serve as an alternative to standard ITT-based $t$ or quantile tests.

### 2.3. Adjustment of discrete covariate

The cQTE estimator can be made more efficient by adjusting for baseline covariates that are predictive of the potential outcomes. With a discrete covariate (e.g. sex and race groups), the adjustment can be done nonparametrically by simple modifications to the procedures described in Section 2.2.

Let $X$ denote the covariate taking values in a discrete set $\mathcal{X}$. Still assuming that $Z$ is completely randomized (so that $Z \perp\!\!\!\perp X$), we can estimate the $\nu_{ac}(\cdot)$ by aggregating the covariate-specific complier outcome distributions. This typically yields a more precise estimator if the covariate is truly associated with the outcome. Specifically, let $\widehat{\nu}_{ac,\mathrm{adj}}(\cdot) = \sum_{x\in\mathcal{X}} n^{-1}n_x\widehat{\nu}_{ac,x}(\cdot)$, where $\widehat{\nu}_{ac,x}(\cdot)$ is the estimator for $\nu_{ac,x}(\cdot) = \mathbb{P}\{Y(a) \leq \cdot \mid A(1) > A(0), X = x\}$ using the subgroup data with $X_i = x$ and $n_x$ is the group size. Then, we can estimate $\mathrm{Q_c}(\tau)$ by $\widehat{\mathrm{Q}}_{\mathrm{c,adj}}(\tau) = \widehat{\nu}^{-1}_{1\mathrm{c,adj}}(\tau) - \widehat{\nu}^{-1}_{0\mathrm{c,adj}}(\tau)$. By derivations only slightly more complicated than those outlined in Section 2.2 (see online supplemental material), we have that $n^{1/2}\{\widehat{\mathrm{Q}}_{\mathrm{c,adj}}(\tau) - \mathrm{Q_c}(\tau)\}$ is asymptotically normal with variance

$$\sigma^2_{\mathrm{adj}}(\tau) = n^{-1}\sum_i^n \left[\eta_{1,X_i}(Z_i, A_i, Y_i)\{\nu^{-1}_{1\mathrm{c}}(\tau)\} - \eta_{0,X_i}(Z_i, A_i, Y_i)\{\nu^{-1}_{0\mathrm{c}}(\tau)\}\right]^2,$$

where $\eta_{a,x}(Z, A, Y)(y) = -\psi_{a,x}(Z, A, Y)(y)/\dot{\nu}_{ac,x}(y)$ $(a = 1, 0)$ and $\psi_{a,x}(Z, A, Y)(y)$ is the influence function for $\widehat{\nu}_{ac,x}(y)$ in the same sense as in (5) (only restricted to the subgroup with $X_i = x$). The functions $\eta_{a,x}(Z, A, Y)(y)$ can be easily estimated by applying the procedures described in Section 2.2 to each subgroup.

### 2.4. Sensitivity analysis

When either (A1) or (A4) is violated, the distributions $\nu_{ac}(y)$ become unidentifiable [30]. Even in such cases, partial information may still exist to allow us to bound the target estimand informatively. First note that $\mathrm{Q_c}(\tau)$ is a monotone functional with respect to the stochastic order of the $\nu_{ac}$ (nondecreasing in $\nu_{1\mathrm{c}}$ and nonincreasing in $\nu_{0\mathrm{c}}$) [33]. If we can find lower and upper stochastic-order bounds for $\nu_{ac}$, denoted by $\underline{\nu}_{ac}$ and $\overline{\nu}_{ac}$, respectively, then the cQTE can be easily bracketed by

$$\mathcal{Q}(\underline{\nu}_{1\mathrm{c}}, \overline{\nu}_{0\mathrm{c}})(\tau) \leq \mathrm{Q_c}(\tau) \leq \mathcal{Q}(\overline{\nu}_{1\mathrm{c}}, \underline{\nu}_{0\mathrm{c}})(\tau), \tag{8}$$

where $\mathcal{Q}(\nu_1, \nu_0)(\tau) = \nu_1^{-1}(\tau) - \nu_0^{-1}(\tau)$ for arbitrary cdf's $\nu_1$ and $\nu_0$.

When exclusion restriction is violated, for example, the target outcome distributions can be redefined as $\nu_{ac}(y) = \mathbb{P}\{Y(a, a) \leq y \mid A(1) > A(0)\}$ $(a = 1, 0)$, where, as described in assumption (A1), $Y(z, a)$ denotes the potential outcome under assigned and received treatments $z$ and $a$, respectively. Let $p_A = \mathbb{P}\{A(1) = A(0) = 1\}$, $p_C = \mathbb{P}\{A(1) > A(0)\}$, and $p_N = \mathbb{P}\{A(1) = A(0) = 0\}$. Under assumptions (A2)–(A4), it can be shown that these compliance class probabilities are still identifiable, through

$$p_A = \mathbb{P}(A = 1 \mid Z = 0), \quad p_N = \mathbb{P}(A = 0 \mid Z = 1), \quad \text{and} \quad p_C = 1 - p_A - p_N.$$

However, the $\nu_{ac}$ are no longer identified. The only identifiable constraint on $\nu_{ac}$ is that it is a component to the distribution in the 'per protocol' group $\mu_a(\cdot) := \mathbb{P}(Y \leq \cdot \mid Z = A =$

$a$), which is a mixture of compliers and noncompliers (always-takers for $\mu_1$ and never-takers for $\mu_0$), with identifiable proportions. In fact, it is easily seen that

$$\nu_{1c} \in \mathcal{P}\left(\frac{p_C}{p_A + p_C}\middle| \mu_1\right) \quad \text{and} \quad \nu_{0c} \in \mathcal{P}\left(\frac{p_C}{p_N + p_C}\middle| \mu_0\right), \tag{9}$$

where $\mathcal{P}(p \mid \mu) = \{\nu : \nu \text{ is a cdf and satisfies } \mu = p\nu + (1-p)\zeta \text{ for some cdf } \zeta\}$. By Proposition 4.3 of Manski [33], $\mathcal{P}(p \mid \mu)$ contains a least and greatest element by stochastic order, which are (quite intuitively) the right- and left-truncated $\mu$ at $\mu^{-1}(p)$ and $\mu^{-1}(1-p)$, i.e.

$$\mathcal{P}^{\min}(p \mid \mu)(\cdot) = \min\{\mu(\cdot)/p, 1\} \quad \text{and} \quad \mathcal{P}^{\max}(p \mid \mu)(\cdot) = 1 - \min[\{1-\mu(\cdot)\}/p, 1], \tag{10}$$

respectively. This is similar to the 'mixture data' approach used by Huber and Mellace [26], Blanco et al. [34,35], and Imai [23] for various purposes. By (10), we obtain that $\underline{\nu}_{1c} = \mathcal{P}^{\min}(\frac{p_C}{p_A+p_C} \mid \mu_1)$, $\overline{\nu}_{1c} = \mathcal{P}^{\max}(\frac{p_C}{p_A+p_C} \mid \mu_1)$, $\underline{\nu}_{0c} = \mathcal{P}^{\min}(\frac{p_C}{p_N+p_C} \mid \mu_0)$, and $\overline{\nu}_{0c} = \mathcal{P}^{\max}(\frac{p_C}{p_N+p_C} \mid \mu_0)$. Replacing the unknown (but identifiable) quantities with their empirical analogs, we can estimate the bounding distributions and construct the bounds for $Q_c(\tau)$ by (8). The resulting bounds are sharp by Sections 4.3 and 4.4 of Manski [33].

**Remark 2.2:** The cQTE considered here combines the effects of both randomization and treatment. To focus on the treatment effect only, Flores and Flores-Lagunes [36] derived bounds for $E\{Y(1,1) - Y(1,0) \mid A(1) > A(0)\}$ under additional monotonicity assumptions on the means of each compliance class. It is possible to use similar stochastic versions of such assumptions to bound $\mathcal{Q}(\nu_{z1,c}, \nu_{z0,c})(\tau)$ $(z = 1, 0)$, where $\nu_{za,c}(y) = \mathbb{P}\{Y(z,a) \leq y\}$ $(a = 1, 0)$. Some discussions are provided in the supplemental material.

In the presence of defiers, the constraints in (9) remain true, but $p_C$, $p_A$, and $p_N$ are no longer point-identified. Under the assumption that $p_C \geq p_D$ [11], however, we can show that the empirical estimators for the mixing proportions $p_C(p_A + p_C)^{-1}$ and $p_C(p_N + p_C)^{-1}$ always underestimate the target proportions. This means that the empirical versions of $\mathcal{P}(\frac{p_C}{p_A+p_C} \mid \mu_1)$ and $\mathcal{P}(\frac{p_C}{p_N+p_C} \mid \mu_1)$ still contain the true ranges of $\nu_{1c}$ and $\nu_{0c}$ (since $\mathcal{P}(p_2 \mid \mu) \subset \mathcal{P}(p_1 \mid \mu)$ for $p_1 \leq p_2$), respectively. Hence, the estimated stochastic-order bounds for the outcome distributions are conservative (i.e. non-sharp) but still valid. To see this, denote $p_A^* = \mathbb{P}(A = 1 \mid Z = 0)$, $p_N^* = \mathbb{P}(A = 0 \mid Z = 1)$, and $p_C^* = 1 - p_A^* - p_N^*$; these are the estimands of the original estimators for $p_A$, $p_N$, and $p_C$, respectively. Let $p_D = \mathbb{P}\{A(1) = 0, A(0) = 1\}$ denote the proportion of defiers. It is not hard to find that $p_A^* = p_A + p_D$, $p_N^* = p_N + p_D$, and $p_C^* = p_C - p_D$, which, by $p_C \geq p_D$, leads to $0 \leq p_C^*(p_A^* + p_C^*)^{-1} \leq p_C(p_A + p_C)^{-1}$ and $0 \leq p_C^*(p_N^* + p_C^*)^{-1} \leq p_C(p_N + p_C)^{-1}$. Hence, the same estimated bounds can be used to bracket the true cQTE when instrument monotonicity is violated.

## 3. Simulation studies

We conducted simulations first to assess the estimation, inference, and sensitivity analysis of the cQTE and then to compare the associated IV test with standard ITT tests for treatment effect.

**Table 1.** Simulation results on the estimation and inference of $Q_c(0.25)$.

| | | High compliance ($p_C = 0.70$) | | | | | Low compliance ($p_C = 0.45$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ITT | cQTE | | | | ITT | cQTE | | | |
| $n$ | $\theta$ | Bias | Bias | SE | SEE | CP | Bias | Bias | SE | SEE | CP |
| 200 | 0 | 0.000 | −0.012 | 0.212 | 0.213 | 0.949 | −0.004 | −0.012 | 0.265 | 0.252 | 0.940 |
| | 0.25 | −0.058 | −0.011 | 0.211 | 0.214 | 0.949 | −0.137 | −0.022 | 0.270 | 0.254 | 0.941 |
| | 0.50 | −0.121 | −0.007 | 0.209 | 0.214 | 0.953 | −0.258 | −0.024 | 0.273 | 0.264 | 0.942 |
| | 1.00 | −0.247 | −0.008 | 0.214 | 0.214 | 0.946 | −0.483 | −0.031 | 0.263 | 0.254 | 0.943 |
| 500 | 0 | −0.003 | 0.005 | 0.136 | 0.135 | 0.944 | 0.002 | 0.011 | 0.146 | 0.143 | 0.947 |
| | 0.25 | −0.056 | −0.006 | 0.131 | 0.134 | 0.952 | −0.139 | −0.011 | 0.148 | 0.144 | 0.952 |
| | 0.50 | −0.118 | 0.005 | 0.132 | 0.134 | 0.949 | −0.257 | −0.009 | 0.147 | 0.143 | 0.951 |
| | 1.00 | −0.244 | −0.005 | 0.133 | 0.134 | 0.950 | −0.480 | −0.012 | 0.148 | 0.143 | 0.949 |
| 1000 | 0 | 0.000 | 0.001 | 0.094 | 0.094 | 0.950 | 0.004 | 0.006 | 0.102 | 0.100 | 0.947 |
| | 0.25 | −0.056 | −0.002 | 0.094 | 0.094 | 0.950 | −0.139 | −0.005 | 0.100 | 0.099 | 0.950 |
| | 0.50 | −0.117 | −0.001 | 0.093 | 0.094 | 0.952 | −0.260 | −0.007 | 0.099 | 0.100 | 0.955 |
| | 1.00 | −0.244 | −0.003 | 0.093 | 0.094 | 0.950 | −0.480 | −0.006 | 0.098 | 0.100 | 0.951 |
| 2000 | 0 | 0.000 | 0.001 | 0.066 | 0.066 | 0.951 | 0.000 | 0.000 | 0.069 | 0.070 | 0.952 |
| | 0.25 | −0.056 | −0.001 | 0.065 | 0.066 | 0.955 | −0.139 | −0.004 | 0.069 | 0.070 | 0.951 |
| | 0.50 | −0.118 | 0.001 | 0.066 | 0.066 | 0.952 | −0.259 | 0.002 | 0.069 | 0.070 | 0.954 |
| | 1.00 | −0.244 | −0.001 | 0.066 | 0.066 | 0.951 | −0.478 | −0.002 | 0.070 | 0.070 | 0.951 |

[a] SE, empirical standard error of the estimator; SEE, empirical average of the standard error estimator; CP, empirical coverage rate of the 95% confidence interval. Each entry is based on 10,000 replicates.

### 3.1. *Estimation of local treatment effects*

Let $Y(0) \sim N(0,1)$ and $Y(1) = Y(0) + \theta$, where $\theta > 0$ denotes the (homogeneous) treatment effect. The compliance status was generated by a conditional trinomial distribution with $\mathbb{P}\{A(1) > A(0) \mid Y(0) = y\} = \exp(-\lambda y^2)$, $\mathbb{P}\{A(1) = A(0) = 1 \mid Y(0) = y\} = \{1 - \exp(-\lambda y^2)\}I(y \leq 0)$, and $\mathbb{P}\{A(1) = A(0) = 0 \mid Y(0) = y\} = \{1 - \exp(-\lambda y^2)\}I(y > 0)$, where $\lambda > 0$ controls the rate of noncompliance. Hence, subjects with 'baseline' outcomes around the mode are most likely compliers; the probability of noncompliance increases as they move away. (This model mimics the situation where, for example, subjects with poorer/better baseline conditions are respectively more/less likely to take the active treatment regardless of the assignment.) Under this set-up, it can be shown that the compliance rate is $p_C = (1 + 2\lambda)^{-1/2}$ and that the complier quantiles are $\nu_{0c}^{-1}(\tau) = (1 + 2\lambda)^{-1/2}\Phi^{-1}(\tau)$ and $\nu_{1c}^{-1}(\tau) = (1 + 2\lambda)^{-1/2}\Phi^{-1}(\tau) + \theta$ (see online supplemental material). We considered two scenarios with $\lambda = 0.5$ and 2, corresponding to compliance rates $p_C = 0.70$ and 0.45, respectively. The results for the estimation and inference of $Q_c(0.25)$ under different values of $\theta$ are summarized in Table 1. For comparison, we also considered the estimator for the ITT effect $Q_{\mathrm{ITT}}(0.25)$. It can be seen that the bias of the ITT estimator with nonzero $\theta$ is substantial, especially when the compliance rate is low. In contrast, the IV estimator $\widehat{Q}_c(0.25)$ exhibits minimal bias across all scenarios. It is worth noting that the variance estimator based on the kernel density estimates with bandwidth specified by the Silverman rule of thumb is consistently accurate, giving rise to 95% confidence intervals with empirical coverage probabilities uniformly close to the nominal rate. Similar results for the estimation and inference of $Q_{\mathrm{ITT}}(0.5)$ and $Q_{\mathrm{ITT}}(0.75)$ are tabulated in the online supplemental material.

Then we assessed the accuracy of the empirical estimator $\widehat{\nu}_0^{-1}(\cdot)$ for the baseline quantile process $\widehat{\nu}_{0c}^{-1}(\cdot)$. The set-up was the same as above except that we fixed $\theta = 0.5$. The

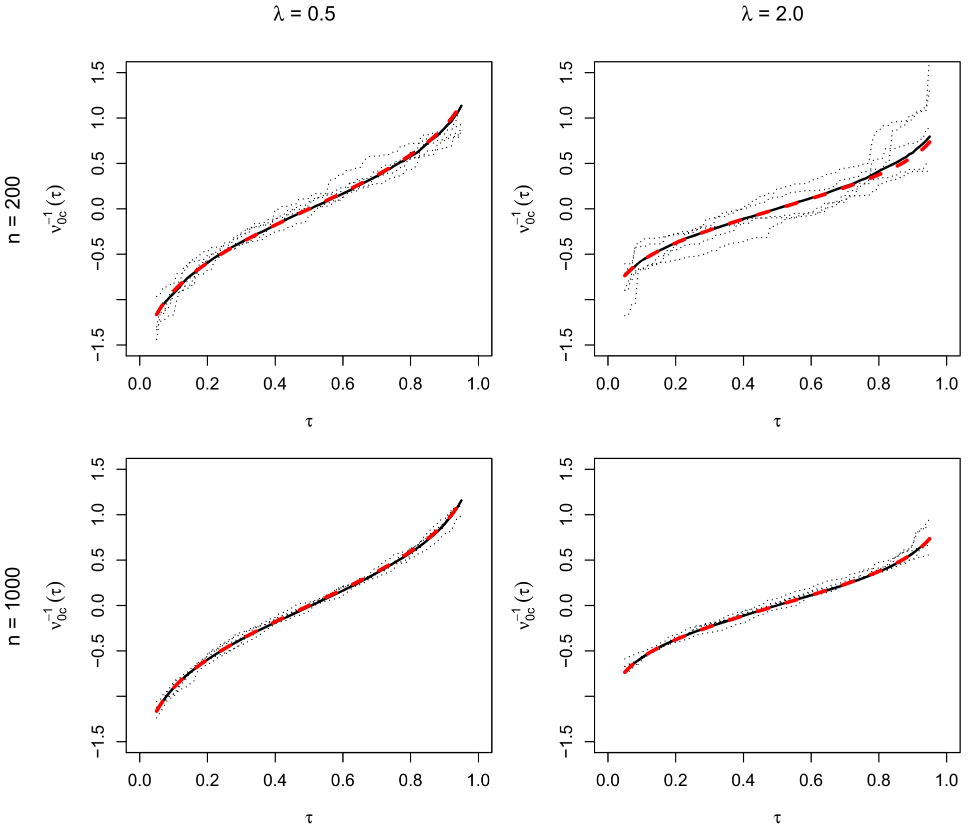

**Figure 1.** Nonparametric estimation of the baseline quantile process $\nu_{0c}^{-1}(\cdot)$. Red dashed line, true function; solid line, average estimates based on 1,000 replicates; dotted lines, five random realizations of the estimated curves.

estimation results are plotted in Figure 1. We can see that the empirical estimator is virtually unbiased for sample size as small as $n = 200$. The estimates become more precise (i.e., with less sampling variation) when $n$ increases to 1000.

Next, we evaluated the accuracy of the sensitivity bounds developed in Section 2.4 in the absence of exclusion restriction. The set-up is the same as before except that the potential outcomes are generated by $Y(z, a) = Y(0, 0) + z\theta/2 + a\theta/2$, so that randomization has additive direct effect $\theta/2$ on all subjects. Under this set-up, expressions for the true bounds given in (8) are derived in the supplemental material. With $\lambda = 0.2, 0.5$ (corresponding to $p_C = 0.85, 0.70$, respectively) and $\theta = 0.5$, we used the 'truncated distributions' in (10) to estimate the upper and lower bounds. With $n = 1000$, the results are displayed in Figure 2. The empirical bounds are seen to have minimal bias with regard to the true bounds.

### 3.2. *Comparison with covariate-adjusted estimator*

In the presence of a discrete covariate, we compared the performance of the covariate-adjusted estimator for the cQTE, as described in Section 2.3, with that of the unadjusted

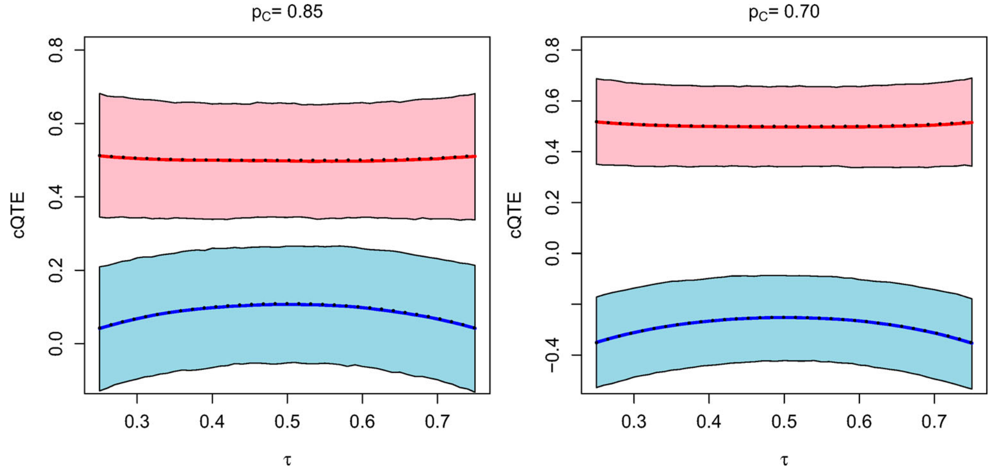

**Figure 2.** Estimation of sensitivity bounds for $Q_c(\tau)$ with $n = 1000$ when exclusion restriction is violated. Dotted lines, true bounds; solid red/blue lines, mean estimates of the upper/lower bounds; shaded areas, range of 95% of the bound estimates. The results are based on 5000 replicates.

**Table 2.** Simulation results comparing the unadjusted versus covariate-adjusted estimators for $Q_c(0.25)$.

| | | | Unadjusted | | | | Adjusted | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma$ | $n$ | $\theta$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP | RE |
| 0 | 200 | 0 | 0.005 | 0.214 | 0.218 | 0.955 | 0.006 | 0.214 | 0.216 | 0.951 | 1.00 |
| | | 0.25 | −0.006 | 0.211 | 0.217 | 0.951 | −0.009 | 0.213 | 0.215 | 0.949 | 0.97 |
| | | 0.5 | −0.011 | 0.213 | 0.218 | 0.955 | −0.012 | 0.215 | 0.216 | 0.949 | 0.98 |
| | | 1 | −0.012 | 0.214 | 0.218 | 0.950 | −0.015 | 0.218 | 0.215 | 0.944 | 0.98 |
| | 500 | 0 | 0.004 | 0.133 | 0.136 | 0.955 | 0.001 | 0.133 | 0.135 | 0.954 | 1.00 |
| | | 0.25 | −0.007 | 0.132 | 0.136 | 0.955 | −0.005 | 0.133 | 0.135 | 0.952 | 0.98 |
| | | 0.5 | −0.003 | 0.133 | 0.135 | 0.955 | 0.001 | 0.134 | 0.135 | 0.952 | 0.99 |
| | | 1 | −0.006 | 0.131 | 0.136 | 0.956 | −0.005 | 0.132 | 0.136 | 0.954 | 0.99 |
| 0.5 | 200 | 0 | 0.014 | 0.247 | 0.251 | 0.95 | 0.003 | 0.232 | 0.227 | 0.942 | 1.13 |
| | | 0.25 | −0.002 | 0.245 | 0.250 | 0.953 | −0.010 | 0.230 | 0.227 | 0.937 | 1.13 |
| | | 0.5 | −0.007 | 0.246 | 0.249 | 0.959 | −0.007 | 0.230 | 0.226 | 0.941 | 1.13 |
| | | 1 | 0.002 | 0.246 | 0.250 | 0.957 | 0.002 | 0.231 | 0.227 | 0.944 | 1.13 |
| | 500 | 0 | −0.002 | 0.149 | 0.155 | 0.957 | −0.006 | 0.140 | 0.142 | 0.954 | 1.15 |
| | | 0.25 | −0.007 | 0.153 | 0.155 | 0.953 | −0.003 | 0.143 | 0.142 | 0.946 | 1.14 |
| | | 0.5 | −0.004 | 0.150 | 0.154 | 0.956 | −0.006 | 0.141 | 0.142 | 0.943 | 1.14 |
| | | 1 | 0.002 | 0.153 | 0.155 | 0.951 | 0.001 | 0.143 | 0.143 | 0.945 | 1.14 |

[a] See note to Table 2. RE, relative efficiency (inverse ratio of variance) of covariate-adjusted versus unadjusted. Each entry is based on 10,000 replicates.

version. Let covariate $X = 1$ and $-1$ with equal probability and $Y(0) \mid Z \sim N(\gamma Z, 1)$, where $\gamma = 0$ and 0.5. The rest of the data generating mechanism is the same as that of Table 1 with $\lambda = 0.5$. Results for the estimation of $Q_c(0.25)$ using both $\widehat{Q}_c(0.25)$ and $\widehat{Q}_{c,\mathrm{adj}}(0.25)$ are summarized in Table 2. Both the adjusted and unadjusted estimators perform satisfactorily in terms of bias, standard error, and confidence interval estimation. When the covariate is independent of the outcome ($\gamma = 0$), the adjusted estimator is no superior to the unadjusted version. In case of a strong covariate-outcome association ($\gamma = 0.5$), adjusting for it increases the efficiency substantially, by 13%–15%.

### 3.3. Power comparison with ITT tests

Finally, we compared the power of the IV test based on $\widehat{Q}_c(0.5)$ (as described in the last paragraph of Section 2.2) with that of the standard ITT median and $t$ tests. For the ITT median test, in particular, we used similar kernel density estimators for the variance of the test statistic. Because of the non-linearity of the median functional, the relationship between the ITT and complier median treatment effects is more complex than that between their ATE counterparts (where the complier effect is famously the ITT effect divided by $p_C$ [11]). As a result, their relative performance in hypothesis testing is unclear and warrants some investigation.

We adopted the same set-up with homogeneous treatment effect $\theta$ as in Section 3.1. For generality, we also considered an additional scenario with inhomogeneous effect. In this latter scenario, we let $Y(1) + 1.5I\{Y(0) < 0\}\theta$. That is, the treatment increases the outcome by $1.5\theta$ if and only if the baseline value is negative. Simulation results on the empirical power of the three tests at level 0.05 are summarized in Table 3. All tests maintain approximately correct type I error (as shown in the empirical rejection rate at $\theta = 0$). The empirical power of the three tests are comparable under homogeneous effect and high compliance rate. However, the $t$ test loses power dramatically under low compliance rate or heterogeneous treatment effect. The IV median test performs similarly to the ITT median test under homogeneous effect, but shows a slight advantage over the latter under inhomogeneous effect, especially when compliance rate is low. In sum, it appears that the median tests are more robust than the $t$ test and that the IV median test is slightly more efficient than the ITT counterpart in cases with low compliance and inhomogeneous effect.

**Table 3.** Simulation results on the empirical power of IV and ITT tests.

| | | Homogeneous effect | | | | | | Inhomogeneous effect | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | High compliance | | | Low compliance | | | High compliance | | | Low compliance | | |
| $n$ | $\theta$ | IV-m | ITT-m | ITT-$t$ | IV-m | ITT-m | ITT-$t$ | IV-m | ITT-m | ITT-$t$ | IV-m | ITT-m | ITT-$t$ |
| 200 | 0 | 0.04 | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 |
| | 0.20 | 0.18 | 0.19 | 0.18 | 0.18 | 0.19 | 0.10 | 0.14 | 0.12 | 0.13 | 0.16 | 0.12 | 0.09 |
| | 0.40 | 0.56 | 0.58 | 0.57 | 0.55 | 0.59 | 0.31 | 0.52 | 0.51 | 0.43 | 0.49 | 0.45 | 0.23 |
| | 0.60 | 0.89 | 0.90 | 0.91 | 0.87 | 0.88 | 0.62 | 0.83 | 0.82 | 0.80 | 0.74 | 0.67 | 0.48 |
| | 0.80 | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 0.89 | 0.96 | 0.96 | 0.98 | 0.85 | 0.80 | 0.79 |
| 500 | 0 | 0.05 | 0.04 | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.04 | 0.04 | 0.05 |
| | 0.10 | 0.13 | 0.13 | 0.13 | 0.13 | 0.14 | 0.09 | 0.07 | 0.07 | 0.10 | 0.09 | 0.07 | 0.06 |
| | 0.20 | 0.39 | 0.41 | 0.37 | 0.39 | 0.42 | 0.19 | 0.33 | 0.31 | 0.25 | 0.35 | 0.28 | 0.13 |
| | 0.30 | 0.73 | 0.75 | 0.70 | 0.72 | 0.75 | 0.37 | 0.67 | 0.66 | 0.52 | 0.66 | 0.60 | 0.26 |
| | 0.40 | 0.93 | 0.94 | 0.92 | 0.92 | 0.93 | 0.62 | 0.90 | 0.89 | 0.79 | 0.88 | 0.84 | 0.46 |
| 1000 | 0 | 0.05 | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 | 0.05 | 0.04 | 0.06 | 0.05 | 0.05 | 0.06 |
| | 0.06 | 0.10 | 0.11 | 0.11 | 0.11 | 0.11 | 0.07 | 0.06 | 0.05 | 0.09 | 0.06 | 0.05 | 0.06 |
| | 0.12 | 0.30 | 0.31 | 0.27 | 0.30 | 0.32 | 0.14 | 0.22 | 0.21 | 0.19 | 0.25 | 0.18 | 0.10 |
| | 0.18 | 0.59 | 0.6 | 0.55 | 0.58 | 0.61 | 0.27 | 0.51 | 0.48 | 0.37 | 0.55 | 0.46 | 0.18 |
| | 0.24 | 0.84 | 0.85 | 0.80 | 0.83 | 0.86 | 0.46 | 0.79 | 0.78 | 0.61 | 0.80 | 0.74 | 0.30 |
| 2000 | 0 | 0.04 | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.04 | 0.04 | 0.05 | 0.05 | 0.05 |
| | 0.04 | 0.09 | 0.10 | 0.10 | 0.11 | 0.12 | 0.08 | 0.06 | 0.05 | 0.08 | 0.06 | 0.05 | 0.07 |
| | 0.08 | 0.27 | 0.28 | 0.25 | 0.28 | 0.3 | 0.13 | 0.19 | 0.17 | 0.18 | 0.21 | 0.16 | 0.10 |
| | 0.12 | 0.55 | 0.56 | 0.50 | 0.55 | 0.57 | 0.24 | 0.44 | 0.42 | 0.33 | 0.50 | 0.43 | 0.16 |
| | 0.16 | 0.80 | 0.81 | 0.74 | 0.80 | 0.82 | 0.39 | 0.72 | 0.70 | 0.52 | 0.75 | 0.68 | 0.26 |

[a] IV-m, test based on $\widehat{Q}_c(0.5)$; ITT-m, median test following the ITT principle; ITT-$t$, $t$ test following the ITT principle. Each entry is based on 10,000 replicates.

## 4. A health insurance program study

We apply our methodology to the Indian health insurance program study mentioned in Section 1. The Rashtriya Swasthya Bima Yojana (RSBY; or National Health Insurance Program) was introduced by the Indian government in 2008 to provide health insurance coverage to the country's low-income residents. A randomized controlled trial was conducted to determine whether enrollment in the Program increases access to hospitalization and health care, as measured primarily by the annual household hospital expenditure [8]. The original study adopted a two-stage randomization scheme, where villages were first randomized to 'high' or 'low' treatment-assignment mechanisms and the households within the village were then randomized to the treatment with 80% or 40% probabilities, respectively. The whole study has recently been analyzed by Imai et al. [8] based on the average treatment effects while accounting for both noncompliance and possible 'spill-over' effects between households in the same village.

For simplicity, we focus on the households under the low treatment-assignment mechanism. Among the 4854 households, 1938 (39.9%) were randomized to the treatment group with free RSBY enrollment and the remaining 2916 (60.1%) were randomized to the control. In the assigned treatment group, 480 (24.8%) households did not enroll in the program, while in the assigned control group, 872 (29.9%) households managed to enroll by other means. The estimated compliance rate is thus $p_C = 1 - 24.8\% - 29.9\% = 45.3\%$. The median annual household hospital expenditures in the assigned treatment and control groups are \$70.2 and \$60.6, respectively. Histograms of the outcomes by randomization and enrollment status are plotted in Figure 3. Given the extreme skewness in the distributions, it is likely more advantageous to analyze the data through the quantiles rather than the average.

Using the procedures described in Section 2.2, we estimated the cQTE $Q_c(\tau)$ for $\tau = 0.05, 0.15, 0.25, 0.5, 0.75, 0.85,$ and $0.95$ along with their 95% confidence intervals using both the unadjusted estimator and one adjusted for the district of the household (Gulbarga versus Mysore). The results are plotted in Figure 4. For comparison, we also indicate the estimated ITT effects in the left panel, which are seen to be smaller than or equal to the corresponding cQTE. The two estimators yield largely similar results, with only slightly narrower confidence interval under the adjusted approach. Overall, the cQTEs appear substantial and significant (at the 0.05 level) at the 0.75- and 0.85-quantiles, while those at or below the median are marginal and nonsignificant. (By the unadjusted estimator, for example, Program enrollment adds \$104.4 to the baseline 0.85-quantile of \$6,190, a 1.7% increase, in the compliers.) There are several plausible explanations for the differential treatment effects. For example, households that spend less on hospital expenditure may tend to be healthier and thus do not stand to gain as much from the government subsidy. It is also possible, however, that there are insurmountable barriers (e.g. financial, cultural, religious, and etc.) to utilization of hospital resources for these households which the RSBY is unable to eliminate or reduce. In any case, this inhomogeneous pattern of treatment effects may signify important features about the causal mechanisms of the RSBY that await further investigation.

**Remark 4.1:** As the original analysis shows by comparing households randomized to 'high' versus 'low' treatment-assignment mechanisms, interferences between households

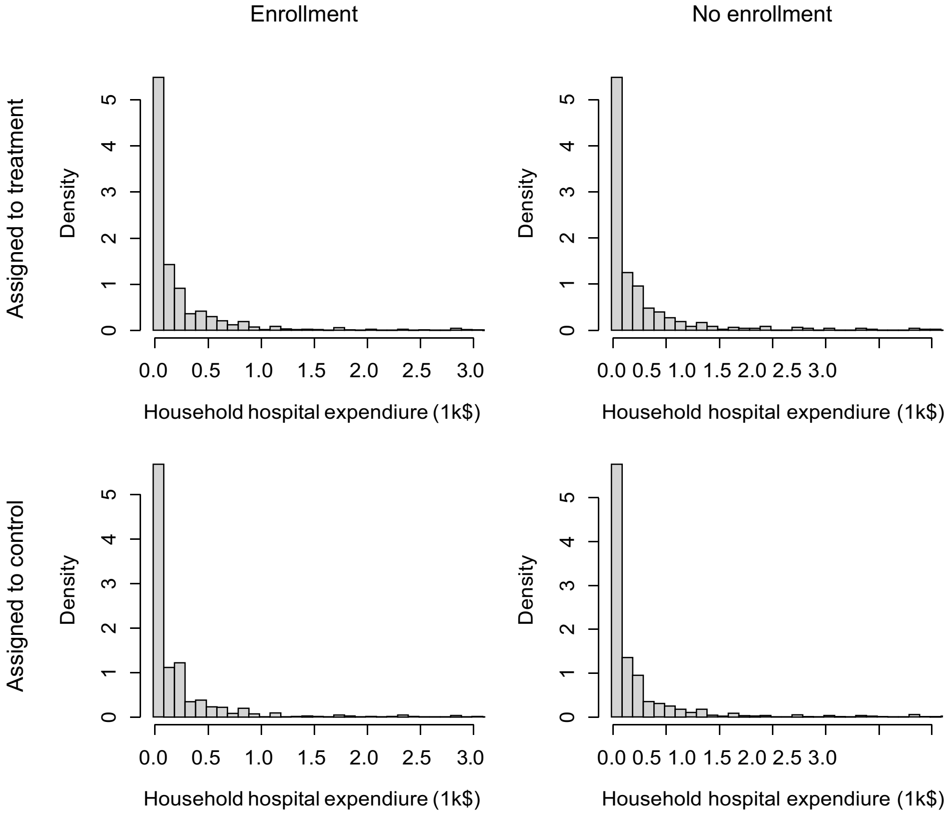

**Figure 3.** Histograms of the outcomes by randomization and enrollment status.

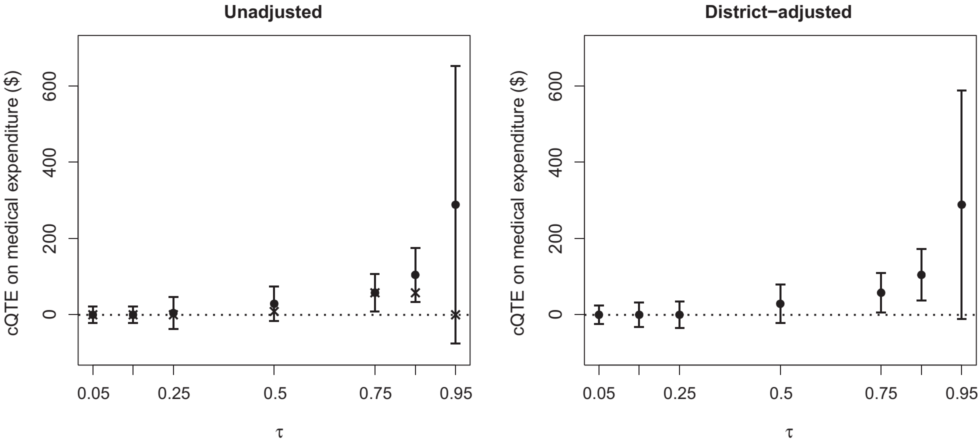

**Figure 4.** Nonparametric estimates for the cQTE (solid circles) and their 95% confidence intervals (error bars) in the RSBY study. Cross sign, estimated $Q_{\mathrm{ITT}}(\tau)$.

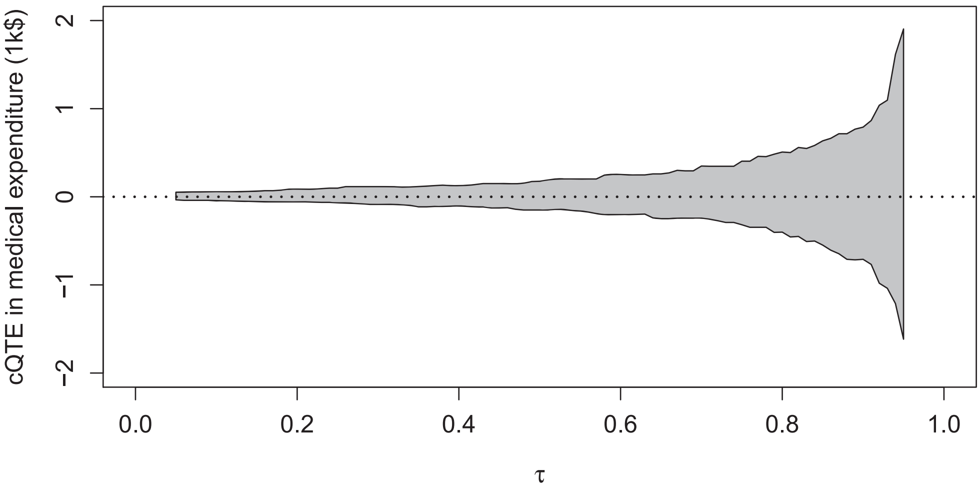

**Figure 5.** Range the partially identified cQTE under violations of exclusion restriction and/or monotonicity in the RSBY study.

within the same village are non-negligible [8]. For example, a household is less likely to utilize hospital resources if neighboring households are also assigned to the Program, possibly due to overcrowded local hospitals. Without considering such spill-over effects, our analysis can overestimate the efficacy of the Program if it is to be rolled out universally.

As argued in Imai et al. [8], both exclusion restriction and monotonicity are plausible assumptions for this study. For illustration, we nonetheless apply the sensitivity analysis techniques described in Section 2.4 to bound the cQTE when these assumptions are violated. The estimated range of the partially identified cQTE under violations of exclusion restriction and/or monotonicity is plotted in Figure 5. Due to the low compliance rate ($p_C = 45.3\%$), however, the bounds are rather wide, especially for the upper quantiles.

## 5. Discussion

Simplifying the general methods of Abadie et al. [15,16] and Frölich and Melly [17], we have studied the complier quantile treatment effect (cQTE) in the special case of randomized trials with noncompliance in a completely nonparametric setting. As shown in the RSBY example, analysis of the quantiles presents a fuller picture of the possibly heterogeneous treatment effects that may be concealed by the mean difference. It is thus advisable in practice to supplement, if not supplant, standard ATE-based analysis with quantile-based analysis such as the cQTE.

The covariate adjustment approach of Section 2.3 has been proven useful in increasing the efficiency of cQTE inference. It is also uniquely straightforward – one just applies the unadjusted method to each level of the covariate and then obtain a pooled estimate, an example of “standardization” [14]. On the other hand, it is applicable only to a discrete covariate with a small number of levels. In case of continuous covariates, one has to either discretize their values into a small number of groups or resort to more sophisticated methods such as described in Frölich and Melly [17], Ye and Lai [37], and Tsiatis [38].

In addition to estimating the local treatment effect, we have also shown by simulation that the cQTE-based IV test tends to outperform the standard ITT tests in adverse conditions such as low compliance rates and heterogeneous treatment effects. The efficiency gain of the IV test is possibly attributable to its exclusive focus on the compliers, the only sub-population in which the treatment effect shows. However, it is important to note that a comparison between the IV and ITT analyses can be made only when both are used to test the causal effect of the treatment. The ITT approach remains indispensable and irreplaceable whenever interest resides in quantifying the causal effect of randomization (such as when evaluating the effect of public policies).

In the absence of exclusion restriction, we have considered bounds for the cQTE defined as the combined effect of both treatment and randomization. To tease out the treatment effect in this case is not easy, as it involves comparison between different treatments in the same randomized group. This comparison is empirically impossible for compliers, whose treatment is fixed by randomization. To circumvent the resulting nonidentifiability, Flores and Flores-Lagunes [36] introduced additional monotonicity assumptions on the outcome means to bound the complier ATE. Both we (see Section S1.5 of supplemental material) and Blanco et al. [34,35] partially extended the mean-monotonicity assumptions to stochastic versions so as to bound the cQTE. More study in this direction is needed in the future.

Our work relies on instrument monotonicity to identify and make inference of the target estimand. An alternate route to identification of the cQTE is by assuming that the ranks of the outcome are preserved by the treatment. Under this rank-preservation condition, the cQTE is not only the difference in the quantiles of the marginal outcome distributions but also the quantile of the individual-level treatment difference. This line of work is pursued by Chernozhukov and Hansen [39,40]. Despite concerns over the plausibility of such assumptions [3,19], it might prove useful to borrow certain ideas from rank preservation in order to relax the monotonicity assumption.

We have been concerned exclusively with binary instrument and treatment. To broaden the scope of cQTE analysis, it will be of interest to consider multivalued instrument and/or treatment as well [41]. Such extensions need not be straightforward as the definition of compliers and the associated assumptions must change accordingly. These too are worthy problems for future research.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was supported by the U.S. National Science Foundation grant DMS2015526.

## Notes on contributor

*Lu Mao* joined the Department of Biostatistics and Medical Informatics at University of Wisconsin (UW)-Madison as an Assistant Professor after obtaining his doctoral degree in Biostatistics from

UNC Chapel Hill in 2016. His research interests include survival analysis (particularly composite outcomes), causal inference, semiparametric theory, and clinical trials. He is currently the PI of an NIH R01 grant on statistical methodology for composite time-to-event outcomes in cardiovascular trials and an NSF grant on causal inference in randomized trials with noncompliance. Besides methodological studies, he also collaborates with medical researchers in cardiology, radiology, cancer, and health behavioral interventions, where time-to-event and longitudinal data are routinely collected and analyzed.

## ORCID

*Lu Mao* http://orcid.org/0000-0002-8626-9822

## References

[1] Lee MJ. Median treatment effect in randomized trials. J R Statist Soc Ser B (Statist Methodol). 2000;62(3):595–604.

[2] Yu K, Lu Z, Stander J. Quantile regression: applications and current research areas. J R Statist Soc Ser D (Statistician). 2003;52(3):331–350.

[3] Firpo S. Efficient semiparametric estimation of quantile treatment effects. Econometrica. 2007;75(1):259–276.

[4] Firpo S, Fortin NM, Lemieux T. Unconditional quantile regressions. Econometrica. 2009;77(3):953–973.

[5] Koenker R. Galton, edgeworth, frisch, and prospects for quantile regression in econometrics. J Econom. 2000;95(2):347–374.

[6] Cade BS, Noon BR. A gentle introduction to quantile regression for ecologists. Front Ecol Environ. 2003;1(8):412–420.

[7] Peng L, Huang Y. Survival analysis with quantile regression models. J Am Stat Assoc. 2008;103(482):637–649.

[8] Imai K, Jiang Z, Malani A. Causal inference with interference and noncompliance in two-stage randomized experiments. J Am Stat Assoc. 2020;116:632–644.

[9] Rubin DB. More powerful randomization-based p-values in double-blind trials with noncompliance. Stat Med. 1998;17(3):371–385.

[10] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[11] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444–455.

[12] Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc. 2005;100(469):322–331.

[13] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge: Cambridge University Press; 2015.

[14] Hernán MA, Robins JM. Causal inference. Boca Raton: CRC Press; 2020.

[15] Abadie A, Angrist J, Imbens G. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica. 2002;70(1):91–117.

[16] Abadie A. Semiparametric instrumental variable estimation of treatment response models. J Econom. 2003;113(2):231–263.

[17] Frölich M, Melly B. Unconditional quantile treatment effects under endogeneity. J Bus Econ Stat. 2013;31(3):346–357.

[18] Eren O, Ozbeklik S. Who benefits from job corps? A distributional analysis of an active labor market program. J Appl Econom. 2014;29(4):586–611.

[19] Melly B, Wüthrich K. Local quantile treatment effects. Bern; 2016. Discussion Papers 16-05. http://hdl.handle.net/10419/149118.

[20] Huber M, Wüthrich K. Local average and quantile treatment effects under endogeneity: a review. J Econom Methods. 2019 Jan;8(1):20170007.

[21] Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Am Stat Assoc. 1997;92(439):1171–1176.
[22] Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by 'deat'. J Educ Behav Stat. 2003;28(4):353–368.
[23] Imai K. Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'. Stat Probab Lett. 2008;78(2):144–149.
[24] Chiba Y. Bounds on the complier average causal effect in randomized trials with noncompliance. Stat Probab Lett. 2012;82(7):1352–1357.
[25] Richardson A, Hudgens MG, Gilbert PB, et al. Nonparametric bounds and sensitivity analysis of treatment effects. Statist Sci: A Rev J Inst Math Statist. 2014;29(4):596.
[26] Huber M, Mellace G. Testing instrument validity for late identification based on inequality moment constraints. Rev Econom Statist. 2015;97(2):398–411.
[27] Jiang Z, Ding P, Geng Z. Principal causal effect identification and surrogate end point evaluation by multiple trials. J R Statist Soc Ser B: Statistical Methodology. 2016;78:829–848.
[28] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
[29] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58(1): 21–29.
[30] Imbens GW, Rubin DB. Estimating outcome distributions for compliers in instrumental variables models. Rev Econ Stud. 1997;64(4):555–574.
[31] Abadie A. Bootstrap tests for distributional treatment effects in instrumental variable models. J Am Stat Assoc. 2002;97(457):284–292.
[32] Silverman BW. Density estimation for statistics and data analysis. Boca Raton: CRC Press; 1986.
[33] Manski CF. Partial identification of probability distributions. New York: Springer Science & Business Media; 2003.
[34] Blanco G, Flores CA, Flores-Lagunes A. Bounds on average and quantile treatment effects of job corps training on wages. J Human Res. 2013;48(3):659–701.
[35] Blanco G, Chen X, Flores CA, et al. Bounds on average and quantile treatment effects on duration outcomes under censoring, selection, and noncompliance. J Bus Econ Stat. 2020;38(4):901–920.
[36] Flores CA, Flores-Lagunes A. Partial identification of local average treatment effects with an invalid instrument. J Bus Econ Stat. 2013;31(4):534-–545.
[37] Ye J, Lai D. Estimations of treatment effects based on covariate adjusted nonparametric methods. Cogent Math Statist. 2020;7(1):1750878.
[38] Tsiatis A. Semiparametric theory and missing data. New York: Springer Science & Business Media; 2006.
[39] Chernozhukov V, Hansen C. An iv model of quantile treatment effects. Econometrica. 2005;73(1):245–261.
[40] Chernozhukov V, Hansen C. Quantile models with endogeneity. Annu Rev Econ. 2013;5(1):57–81.
[41] Lee S, Salanié B. Identifying effects of multivalued treatments. Econometrica. 2018;86(6):1939–1963.