Statistica Sinica Preprint No: SS-2022-0288

Title	A Data Fusion Method for Quantile Treatment Effects
Manuscript ID	SS-2022-0288
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0288
Complete List of Authors	Yijiao Zhang and Zhongyi Zhu
Corresponding Authors	Zhongyi Zhu
E-mails	zhuzy@fudan.edu.cn

Notice: Accepted version subject to English editing.

A DATA FUSION METHOD FOR QUANTILE TREATMENT EFFECTS

Yijiao Zhang and Zhongyi Zhu

Department of Statistics and Data Science, Fudan University

Abstract: With the increasing availability of datasets, developing data fusion methods to leverage the strengths of different datasets to draw causal effects is of great practical importance to many scientific fields. In this paper, we consider estimating the quantile treatment effects using small validation data with fully-observed confounders and large auxiliary data with unmeasured confounders. We propose a Fused Quantile Treatment effects Estimator (FQTE) by integrating the information from two datasets based on doubly robust estimating functions. We allow for the misspecification of the models on the dataset with unmeasured confounders. Under mild conditions, we show that the proposed FQTE is asymptotically normal and more efficient than the initial QTE estimator using the validation data solely. By establishing the asymptotic linear forms of related estimators, convenient methods for covariance estimation are provided. Simulation studies demonstrate the empirical validity and improved efficiency of our fused estimators. We illustrate the proposed method with an application.

Key words and phrases: Calibration; Causal Inference; Double Robustness; Estimation Equation; Unmeasured Confounder.

1. Introduction

The increasing availability of datasets from multiple sources holds enormous promise for evaluating causal effects. With various datasets at hand, data fusion technology has become more and more important in many medical and biological applications. How to systematically combine multiple datasets sources in an attempt to leverage the strengths of different types of data to improve the estimating efficiency of causal effects is gathering notice from researchers. For example, there are data sources with large sample size, such as electronic health records, claims databases, disease data registries, and census data. However, uncontrolled design mechanisms and limited information on baseline covariates may lead to confounding bias, presenting a major threat to causal inference. In practice, there are also small validation datasets that include all possible confounders and provide detailed information for each individual, especially in some randomized controlled trials (RCTs) in the medical field. A classic example is a two-phase study (Wang et al., 2009), where less expensive covariates are measured for all subjects in the first phase and the detailed information is collected in the second phase only for a validation subset drawn from the full sample. Unfortunately, the validation datasets often suffer from limited sample size due to the limitation of cost. Therefore, evaluating causal effects based solely on the validation datasets lacks efficiency. Consequently, we are seeking estimators of higher efficiency while pursuing unbiasedness as well by integrating information from both types of datasets.

In the literature on causal inference, a great much of attention has been paid to the average treatment effects (ATE). In addition, quantiles are also useful measures for detecting causal effects. Firstly, when the outcomes are distributed with heavy tails, the medians are more efficient than the means. Secondly, quantiles are more appropriate measures when the distributions of outcomes are skewed. Thirdly, quantiles can also provide a more detailed view of heterogeneous causal effects at different points. In particular, researchers or policy-makers may be more interested in the distributional impacts on the dispersion of the outcome or the lower or higher tail of the distributions of potential outcomes beyond the average effects of treatment.

In this article, we consider the data fusion problem of estimating the quantile treatment effects (QTE), defined as the difference between the quantiles of the marginal potential distributions of the treatment and control responses. We focus on the case where there are two types of data sources. One is a validation dataset that includes the measurements of all the confounders but suffers from small sample size and the other is an auxiliary dataset that enjoys large sample size but has unmeasured confounders.

In the case where there are no unmeasured confounders, several works have been done on identifying and estimating the conditional or unconditional QTE; for example, Firpo (2007), Zhang et al. (2012), Donald and Hsu (2014), to name a few. (Firpo, 2007) proposed an inverse probability weighting (IPW) estimator based on a nonparametric power series estimator of the propensity score and showed that under regular conditions their IPW estimator is root n consistent and achieves the semiparametric efficiency bound. (Zhang et al., 2012) proposed an outcome regression (OR) estimator and a parametric inverse probability weighting (IPW) estimator based on a pre-specified outcome model and propensity score model respectively. They also proposed a doubly robust (DR) estimator which is consistent if either the outcome model or the propensity score model is correctly specified. In these papers, the unconfoundedness treatment assignment and strict overlap assumption are assumed. However, in observational studies, the unconfoundedness assumption may be violated due to unmeasured confounders, which is also called endogeneity of the treatment variable in the economics literature. To deal with this, instrumental variable (IV) approaches are developed for identifying the average or quantile treatment effects, see, for example, Imbens and Angrist (1994), Wüthrich

(2019) for details. Nevertheless, valid instrumental variables are often difficult to find in practice.

A burgeoning literature on data fusion has explored the possibility of harmonizing evidence from multiple data sources for estimating causal effects. Refer to Colnet et al. (2020) for a detailed review. When the unconfoundedness is not assumed in the auxiliary big dataset, several methods are developed to deal with the confounding bias. One line is to construct shrinkage estimators by combining unbiased and biased estimators (Rosenman et al., 2022; Cheng and Cai, 2021), which achieve lower MSE than the initial unbiased estimator based solely on the validation data. Another line is to specify a parametric model for the confounding bias (Kallus et al., 2018; Yang et al., 2020). However, the bias model is difficult to be correctly specified, especially in the case of QTE estimation.

Empirical likelihood approaches are also commonly used for integrating information from multiple data sources (Chatterjee et al., 2016; Zhang et al., 2020). Chatterjee et al. (2016) consider a constrained maximum likelihood estimator using summary-level information from an external study, which is shown to be more efficient than the estimator based only on internal sample data. However, their method focuses on the regression parameters and can not be applied directly to the causal inference framework. Besides, the empirical likelihood approach always needs heavy computation when the data size gets large.

Considerably less work is available in this literature for QTE estimation under the data fusion framework. To the best of our knowledge, the only one is by Li and Luedtke (2021), which proposed a general semiparametric framework for efficient estimation under data fusion, including the estimation of QTE. However, their estimation is based on the canonical gradient, which may be sophisticated when there are unmeasured confounders.

Under our framework, though the two types of data may not be combined directly, they may share some common information. A natural idea is to connect the shared common information in multiple data sources through a calibration technique (see e.g. Wu and Sitter, 2001; Lin and Chen, 2014). More related to our work, Yang and Ding (2020) used a calibration idea to improve the efficiency of the initial estimators of ATE by projecting them to the difference between two error-prone estimators based on the validation data and the auxiliary data respectively. A similar idea is also used in Cai et al. (2021) for developing the optimal decision rule. The key insight is that the differences should be consistent estimators for zero. However, due to the essential properties of quantiles, a direct application of the difference-based method for QTE estimation involves the estimation of covariance matrices with rather complicated forms. Bootstrap methods for variance estimation could be time-consuming since the estimators themselves are constructed based on estimated covariance.

In this article, we propose a fused quantile treatment effects estimator (FQTE) by integrating the information from both the validation data and the auxiliary data through estimating functions. We break down our contribution as follows:

On the methodological side, we show how to connect the two datasets through estimating functions and project the initial estimators on them to obtain our FQTE. We rationalize our idea from three different perspectives. Firstly, we can treat the biased estimators for quantiles as summary-level information from the big main data and use the estimation equation of the summary-level information as moment conditions on the validation dataset to make calibration to our initial estimators. Secondly, these estimating functions can be interpreted as linear combinations of doubly robust rank scores, which preserve the main information for quantiles robustly. Thirdly, as these estimating functions are consistent estimators for zero as well, we show our method as a generalization of Yang and Ding (2020) by reformulating the difference-based estimators for zero thereof.

On the theoretical side, there are two core results. Firstly, we establish the asymptotic linear representations of the initial QTE estimators as well as those estimating functions we project on. This is the fundamental property that does not hold if we use the difference-based method in Yang and Ding (2020) for QTE estimation. Thanks to these asymptotic linear representations, convenient covariance estimation methods are provided to make our method easy to implement. To derive the asymptotic linear representations, the non-smooth estimating functions are dealt with via empirical process theory (e.g. Kosorok, 2008, Chap. 8). Secondly, consistency and asymptotic normality of our FQTE are established and we show that FQTEs enjoy efficiency gains. Besides, we also establish the consistency of our variance estimators based on the asymptotic variance.

By applying a missing mechanism to the datasets, we further extend our method to the cases where the validation sample may not be a random sample from the entire dataset, which is more reasonable in practice. Estimators and asymptotic results are provided based on unknown missing probabilities, which are assumed to be observed in Yang and Ding (2020).

The rest of this paper is organized as follows. In Section 2, we give an exposition of the problem setup. We begin in Section 3 by proposing our data fusion method and then provide heuristic explanations. Section 4 establishes the theoretical properties. The finite sample properties on simulated and real datasets are investigated in Section 5 and Section 6. We conclude this paper with a discussion in Section 7.

2. Setup and Basic Estimators

2.1 Basic Notations

We focus on the scenario where there is a validation dataset with fully-observed confounders and an auxiliary dataset with partially-observed confounders. We adopt the potential outcomes framework of Neyman and Rubin. See Rubin (1974). We focus on a binary treatment $T \in \{0, 1\}$, which is an intervention of interest. Let $Y(t), t \in \{0, 1\}$ denote the potential outcomes, which we interpret as the outcome had the individual assigned to treatment t. We assume the consistency assumption always holds, that is, the observed outcome Y with an assigned treatment T equal to t equal to its potential outcome $Y(t)$, i.e., $Y = Y(T) = TY(1) + (1 - T)Y(0)$. Let X denote a p_x-dimensional vector of pre-treatment baseline covariates with support $\mathbb{X}$, S a p_s-dimensional vector of pre-treatment baseline covariates with support $\mathbb{S}$ and R a binary indicator for being in the validation sample or not ($R_i = 1$ if the ith individual is in the validation data and $R_i = 0$ otherwise). We model each individual in the observed data by a random tuple $(Y(1), Y(0), T, X^\top, S^\top, R)$ drawn from a superpopulation $\mathcal{P}$. We use pr to denote the distribution under $\mathcal{P}$ and denote E as the expectation operator under pr. The basic information X is observed for all individuals, but the more detailed information S is observed only on a subset of individuals. Denote the full-observed information as $O = (Y, T, X^\top, S^\top)$ and the partially-observed information as $U = (Y, T, X^\top)$ with support $\mathbb{U}$. The validation dataset $\{O_i = (Y_i, T_i, X_i^\top, S_i^\top) : i = 1, ..., n\}$ consist of n identically and independently distributed (i.i.d.) observations , while the auxiliary data $\{U_i = (Y_i, T_i, X_i^\top) : i = n+1, ..., N\}$ consist of $m = N - n$ i.i.d observations without S. Define $\nu_n = n/N$ as the sample ratio between the validation data and the entire observed data and $\nu_n \to \nu \in [0, 1)$ as $n \to \infty$. The entire observed data could thus be formulated as $\{D_i = (R_i, Y_i, T_i, X_i^\top, R_i S_i^\top) : i = 1, ..., N\}$. Define the index set $\mathcal{V} = \{1, ..., n\}$ and $\mathcal{O} = \{1, ..., N\}$.

Define $F_t(y \mid X, S) = \mathrm{pr}(Y \leq y | T = t, X, S)$ and $F_t(y \mid X) = \mathrm{pr}(Y \leq y \mid T = t, X)$ as the conditional distribution of the observed outcome given the fully-observed covariates (X, S) and partially-observed covariates X respectively. Denote the conditional probability of the treatment as

$$e(X, S) = \mathrm{pr}(T = 1 \mid X, S), \quad e(X) = \mathrm{pr}(T = 1 \mid X).$$

The former is known as the propensity score in the causal inference literature. We call the latter one the pseudo propensity score as it does not include the information of the unmeasured confounders S.

Now we consider the quantile treatment effects. Denote p as the quantile level and $F_t(\cdot)$ as the marginal cumulative distribution function of $Y(t)$. Formally, for any given quantile level $p \in (0, 1)$, the pth quantile treatment effect is defined as

$$\Delta_p = q_{1,p} - q_{0,p},$$

where $q_{t,p} = \inf\{q : F_t(q) \geq p\}$ is the pth quantile of $F_t(\cdot)$.

2.2 Estimators Using the Fully-observed Validation Data

The following are classical assumptions for identifying the quantile treatment effects.

Assumption 1 (Ignorability)**.** $Y(t) \perp\!\!\!\perp T \mid (X, S)$ *for* $t = 0, 1$.

Assumption 2 (Overlap)**.** *There exist constants* c_1 *and* c_2 *such that with probability* $1, 0 < c_1 \leq e(X, S) \leq c_2 < 1$.

The distribution of the potential outcomes can be identified under Assumptions 1 and 2 and classical estimators have been developed for estimating the QTE, including the outcome regression (OR) (Zhang et al., 2012), inverse probability weighting (IPW), doubly robust (DR) estimators (Firpo, 2007; Zhang et al., 2012; Donald and Hsu, 2014). We assume that

Assumptions 1 and 2 hold hereafter. Therefore, we can obtain an initial estimator using the validation data only.

Let $G_t(y \mid X, S; \theta_t)$ be a parametric working outcome regression (OR) model for $F_t(y \mid X, S)$, for example, a normal linear model after a Box-Cox transformation. Let $e(X, S; \alpha)$ be a parametric working propensity score (PS) model for $e(X, S)$. A common choice would be a logistic regression model. Let $\hat{\theta}_t^{\mathcal{V}}$ and $\hat{\alpha}^{\mathcal{V}}$ be consistent estimators for the corresponding true parameters θ_t^* and α^* based on the validation sample, for example, the maximum likelihood estimator (MLE). For simplicity, we omit the superscript $\mathcal{V}$ and denote the estimators as $\hat{\theta}_t$ and $\hat{\alpha}$ hereafter with no ambiguity. Define the weights $w_{1,i}^* = T_i/e\left(X_i, S_i; \alpha^*\right)$, $w_{0,i}^* = (1-T_i)/(1-e\left(X_i, S_i; \alpha^*\right))$ and the estimated weights $\hat{w}_{t,i}$ with α^* in $w_{t,i}^*$ replaced by its estimates $\hat{\alpha}$. Further denote $T/e\left(X, S; \alpha^*\right)$ as w_1^*, and $(1-T)/(1-e(X, S; \alpha^*))$ as w_0^*.

Assumption 3 (Outcome Model)**.** *The parametric model $G_t(y \mid X, S; \theta_t)$ is a correct specification for $F_t(y \mid X, S)$, for $t = 0, 1$, that is, $F_t(y \mid X, S) = G_t(y \mid X, S; \theta_t^*)$, where G is a known function and θ_t^* is the true model parameter, for $t = 0, 1$.*

Similar assumptions about the correct specification of conditional distribution have been proposed in Zhang et al. (2012) and Han et al. (2019) for quantile estimation with missing data. We may relax it to the correct specification of the conditional quantile, which is discussed in Section 7.2.

Assumption 4 (Propensity score model)**.** *The parametric model* $e(X, S; \alpha)$ *is a correct specification for* $e(X, S)$*; that is,* $e(X, S) = e\left(X, S; \alpha^*\right)$*, where* α^* *is the true model parameter.*

By modelling both the conditional distribution $F_t(y \mid X, S)$ and the propensity score $e(X, S)$, we arrive at the so-called doubly robust (DR) estimator (Zhang et al., 2012). For simplicity, let $\eta_t = (\theta_t, \alpha)$ denote the nuisance parameter, with $\eta_t^* = (\theta_t^*, \alpha^*)$ being its true value and $\hat{\eta}_t = (\hat{\theta}_t, \hat{\alpha})$ being its estimator. Under Assumptions 3 or 4, $q_{t,p}$ $(t = 0, 1)$ can be identified by $E\left\{\Psi_t(O; q_{t,p}, \eta_t^*)\right\} = 0$, where

$$\Psi_t(O; q_{t,p}, \eta_t^*) = w_t^* \left\{I(Y \leqslant q_{t,p}) - G_t\left(q_{t,p} \mid X, S; \theta_t^*\right)\right\} + G_t\left(q_{t,p} \mid X, S; \theta_t^*\right) - p. \tag{2.1}$$

A DR estimator for Δ_p based on the validation sample is defined as $\hat{\Delta}_p^{\mathcal{V}}$ $= \hat{q}_{1,p}^{\mathcal{V}}$ $- \hat{q}_{0,p}^{\mathcal{V}}$, where $\hat{q}_{t,p}^{\mathcal{V}}$ is an DR quantile estimator for $q_{t,p}(t = 0, 1)$, which is the solution to

$$1/n \sum_{i=1}^{n} \Psi_t(O_i; q, \hat{\eta}_t) = 0, \tag{2.2}$$

where $\Psi_t(O_i; q, \hat{\eta}_t) = \hat{w}_{t,i}\{I(Y_i \leqslant q) - G_t(q \mid X_i, S_i; \hat{\theta}_t)\} + G_t(q \mid X_i, S_i; \hat{\theta}_t) - p$.

The sum of weights $\sum_{i=1}^{n} \hat{w}_{t,i}$ converges to 1 as $n \to \infty$ but is generally different from 1 for any finite n. For improved finite-sample performance, normalized weights can be calculated. The DR estimator is consistent if either the OR or the PS model is correctly specified. Moreover, the DR estimator is locally efficient when both outcome and propensity score models are correctly specified (Díaz, 2017).

Under regular conditions, the DR estimator is also asymptotically linear in the sense of (Tsiatis, 2007). To be detailed, according to Theorem 1 in Section 4.1, we have for the DR quantile estimators that

$$n^{1/2}(\hat{q}^{\mathcal{V}}_{t,p} - q_{t,p}) = 1/n^{1/2} \sum_{i=1}^{n} \psi_t(O_i; q_{t,p}, \eta^*_t) + o_p(1). \qquad (2.3)$$

where $\psi_t(O_i; q_{t,p}, \eta^*_t)$ is given in (4.12) in Section 4.1 and it is called the influence function of $\hat{q}^{\mathcal{V}}_{t,p}$. The asymptotic linear representations provide us with great convenience for variance estimation.

2.3 Estimators Using the Partially-observed Entire Data

The initial estimators only use information from the validation sample which has a small sample size and hence lacks efficiency. That's why we need the auxiliary datasets with a large sample size to help improve the efficiency. However, the auxiliary dataset does not include detailed information about S and hence may lead to confounding bias. Nevertheless, we may treat X as all the confounders and use the entire data with only $U = (Y, T, X^\top)$ to obtain quantile and QTE estimators, following the same estimating procedure as we do on the validation dataset.

To be specific, we may consider using the same working models as that in Section 2.2 for the conditional distribution $F_t(y \mid X)$ (for example, both normal linear) and the pseudo propensity score $e(X)$ (for example, both logistic), which may be misspecified. Denote them as $\tilde{G}(X;\theta_t^{\text{Conf}})$ and $\tilde{e}(X;\alpha^{\text{Conf}})$ respectively. Here "Conf" is short for "confounded". Let $\hat{\theta}_t^{\text{Conf},\mathcal{O}}$ and $\hat{\alpha}^{\text{Conf},\mathcal{O}}$ be the corresponding MLEs based on the entire sample, with probability limits $\theta_t^{\text{Conf},*}$ and $\alpha^{\text{Conf},*}$ (White, 1982). For simplicity, we omit the superscript "$\mathcal{O}$" and denote the estimators as $\hat{\theta}_t^{\text{Conf}}$ and $\hat{\alpha}^{\text{Conf}}$ hereafter. Define the weights $z^*_{1,i} = T_i/e\left(X_i;\alpha^{\text{Conf},*}\right)$, $z^*_{0,i} = (1-T_i)/(1-e\left(X_i;\alpha^{\text{Conf},*}\right))$ and the estimated weights $\hat{z}_{t,i}$ with $\alpha^{\text{Conf},*}$ in $z^*_{t,i}$ replaced by its estimates $\hat{\alpha}^{\text{Conf}}$. Further denote $T/e\left(X;\alpha^{\text{Conf},*}\right)$ as z^*_1 and $(1-T)/(1-e(X;\alpha^{\text{Conf},*}))$ as z^*_0. Similarly, let $\eta_t^{\text{Conf}} = (\theta_t^{\text{Conf}}, \alpha^{\text{Conf}})$ denote the nuisance parameter, with $\eta_t^{\text{Conf},*} = (\theta_t^{\text{Conf},*}, \alpha^{\text{Conf},*})$ being its probability limit and $\hat{\eta}_t^{\text{Conf}} = (\hat{\theta}_t^{\text{Conf}}, \hat{\alpha}^{\text{Conf}})$ being its consistent estimator.

We may construct empirical estimation equations similar to (2.2) to obtain DR quantile estimators $\hat{q}_{t,p}^{\text{Conf}}$ subject to unmeasured confounding, by solving the equation (in q)

$$1/N\sum_{i=1}^{N}\phi_t(U_i;q,\hat{\eta}_t^{\text{Conf}}) = 0, \tag{2.4}$$

where $\phi_t(U_i;q,\hat{\eta}_t^{\text{Conf}}) = \hat{z}_{t,i}\{I(Y_i \leqslant q) - \tilde{G}_t(q \mid X_i;\hat{\theta}_t^{\text{Conf}})\} + \tilde{G}_t(q \mid X_i;\hat{\theta}_t^{\text{Conf}}) - p$.

In fact, the parameter of which $\hat{q}_{t,p}^{\text{Conf}}$ estimates , denoted as $q_{t,p}^{\text{Conf}}$, can be identified by the following equation on the partially-observed data

$$E\left\{\phi_t(U; q_{t,p}^{\text{Conf}}, \eta_t^{\text{Conf},*})\right\} = 0. \tag{2.5}$$

We call $q_{t,p}^{\text{Conf}}$ as the pseudo quantile. To explain it, if the unconfoundedness assumption also holds with the partially-observed confounders X, we have that $q_{t,p}^{\text{Conf}}$ is equal to $q_{t,p}$ if either $\tilde{G}(X; \theta_t^{\text{Conf}})$ is correctly specified for $F_t(y \mid X)$ or $\tilde{e}(X; \alpha^{\text{Conf}})$ is correctly specified for $e(X)$. Then we will obtain $N^{1/2}$-consistent estimators for the quantiles and then the QTE, which are more efficient than the initial estimators. However, due to the unmeasured confounders S, $q_{t,p}^{\text{Conf}}$ can be different from $q_{t,p}$, which leads $\hat{q}_{t,p}^{\text{Conf}}$ to be biased estimators for our interested parameter $q_{t,p}$. Consequently, the QTE estimators using X only are often inconsistent.

The question is how to integrate the unbiased but inefficient estimator $\hat{q}_{t,p}^{\mathcal{V}}$ in Section 2.2 and the more efficient but biased estimator $\hat{q}_{t,p}^{\text{Conf}}$ here to improve the efficiency of the initial estimators while pursuing consistency as well. This is what we do in the next section.

3. Method

3.1 Proposed Method

Now we consider using the calibration technique to fuse the two datasets to draw inference for the QTE. The key idea is to connect the two datasets through estimating functions and then take projection.

Though the two types of data may not be fused directly, they may share some common information. A natural idea is to connect the shared common information in multiple data sources through a calibration technique (see e.g. Lin and Chen, 2014). To make a calibration, we need assumptions about the shared information on the two datasets. The following assumption is classical in the missing data literature.

Assumption 5 (Missing Completely at Random, MCAR)**.** $R \perp\!\!\!\perp (Y, T, X, S)$.

We further extend this MCAR assumption to a weaker missing at random (MAR) assumption ($R \perp\!\!\!\perp S|(Y, T, X)$) in the supplementary materials Section S2, which allows the selection of the validation sample to depend on a probability design. Under Assumption 5, $\{(Y_i, T_i, X_i^\top), i = 1, ..., N\}$ are i.i.d. samples. Then we have that the equation $E\left\{\phi_t(U; q_{t,p}^{\text{Conf}}, \eta_t^{\text{Conf}})\right\} = 0$ will also hold on the validation sample, which motivates us to connect the two datasets through the estimating functions

$$\hat{C}_t = 1/n \sum_{i=1}^{n} \phi_t(U_i; \hat{q}_{t,p}^{\text{Conf}}, \hat{\eta}_t^{\text{Conf}}), \tag{3.6}$$

where $\phi_t(U_i; \cdot, \cdot)$ is defined in (2.4) in Section 2.3. Note that (3.6) integrates the information from the auxiliary data through $\hat{q}_{t,p}^{\text{Conf}}$ and $\hat{\eta}_t^{\text{Conf}}$ as well as the information from the validation data through $\{U_i\}_{i=1}^n$.

Similar to (2.3), under regular conditions given in detail in Section 4.1, we can also establish the asymptotical linear representations for $\hat{C}_t$ as

$$\hat{C}_t = 1/n \sum_{i=1}^{n} \phi_t(U_i; q_{t,p}^{\text{Conf},*}, \eta_t^{\text{Conf},*}) - 1/N \sum_{i=1}^{N} \phi_t(U_i; q_{t,p}^{\text{Conf},*}, \eta_t^{\text{Conf},*}) + o_p(n^{-1/2}), \tag{3.7}$$

which also implies that $\hat{C}_t$ is a consistent estimator for zero. Here we don't need any assumptions on the correct specification of models on the joint distribution of $U = (A, X, Y)$. We allow both the working models $F(y \mid X)$ and $e(X)$ to be misspecified. For similicity, denote $\phi_t(U_i; q_{t,p}^{\text{Conf},*}, \eta_t^{\text{Conf},*})$ as $\phi_{t,i}$ and $\psi_t(O_i; q_{t,p}, \eta_t^*)$ as $\psi_{t,i}$ for $t = 0, 1$. Combining (2.3) and (3.7), the next proposition models the asymptotic joint distribution of $\hat{\Delta}_p^{\mathcal{V}}$ and $\hat{C}_t$.

Proposition 1. *Under the assumptions in Theorem 1 and Lemma 1 in Section 4.1, as $n \to \infty$, then*

$$n^{1/2} \begin{pmatrix} \hat{\Delta}_p^{\mathcal{V}} - \Delta_p \\ \hat{C} \end{pmatrix} \longrightarrow \mathcal{N} \left\{ 0, \begin{pmatrix} \sigma_{\mathcal{V}}^2 & \varrho^\top \\ \varrho & \Sigma_{\text{ep}} \end{pmatrix} \right\}, \tag{3.8}$$

in distribution, where $\hat{C} = (\hat{C}_1^\top, \hat{C}_0^\top)^\top$*,* $\sigma^2_{\mathcal{V}} = \mathrm{var}(\psi_{1,i} - \psi_{0,i})$*,* $\varrho = (1 - \nu)\,\mathrm{cov}(\psi_{1,i} - \psi_{0,i}, (\phi_{1,i}^\top, \phi_{0,i}^\top)^\top)$*,* $\Sigma_{\mathrm{ep}} = \begin{pmatrix} \Sigma_1 & \Sigma_{01}^\top \\ \Sigma_{01} & \Sigma_0 \end{pmatrix}$ *with* $\Sigma_{01} = (1 - \nu)\,\mathrm{cov}(\phi_{0,i}, \phi_{1,i})$ *and* $\Sigma_t = (1 - \nu)\mathrm{var}(\phi_{t,i})$*, for* $t = 0, 1$*.*

Remark 1. Proposition 1 is analogous to Theorem 1 in Yang and Ding (2020) for estimating ATE, which is the fundamental part of our method.

By projecting $\hat{\Delta}_p^{\mathcal{V}}$ to $\hat{C}$, we can obtain our fused QTE estimator (FQTE)

$$\hat{\Delta}_p = \hat{\Delta}_p^{\mathcal{V}} - \hat{\varrho}^\top \hat{\Sigma}_{\mathrm{ep}}^{-1} \hat{C}, \tag{3.9}$$

where $\hat{\varrho}$ and $\hat{\Sigma}_{\mathrm{ep}}$ are corresponding consistent estimators for ϱ and Σ_{ep}. The construction of consistent covariance estimators $(\hat{\varrho}, \hat{\Sigma}_{\mathrm{ep}})$ in (3.9) will be discussed in the Section 4.3. We assume that Σ_{ep} is positive definite, which is similarly assumed in Yang and Ding (2020).

Remark 2. We can also integrate the information of pseudo quantiles at different orders to draw inference about the QTE at the pth order. Consider a d-dimensional vector q_t^{Conf}, $t = 0, 1$, with the kth dimension be the p_kth $(k = 1, \ldots, d)$ order pseudo quantile $q_{t,p_k}^{\mathrm{Conf}}$ identified by the equations $E\phi_{t,k}(U_i; q_{t,p_k}^{\mathrm{Conf}}, \eta_t^{\mathrm{Conf},*}) = 0, t = 0, 1$, where $\phi_{t,k}(U_i; q_{t,p_k}^{\mathrm{Conf}}, \eta_t^{\mathrm{Conf},*})$ equals

$$z_{t,i}^* \left\{ I(Y_i \leqslant q_{t,p_k}^{\mathrm{Conf}}) - \tilde{G}_t(q_{t,p_k}^{\mathrm{Conf}} \mid X_i; \theta_t^{\mathrm{Conf},*}) \right\} + \tilde{G}_t(q_{t,p_k}^{\mathrm{Conf}} \mid X_i; \theta_t^{\mathrm{Conf},*}) - p_k$$

Then we will obtain two d-dimensional vector $\hat{C}_1$ and $\hat{C}_0$. We denote hereafter the pseudo quantiles $(p_1, p_2, \dots, p_d)$ chosen for calibration as p_{cal}.

3.2 Heuristic explanation

The construction of $\hat{C}_t$ is motivated by the usage of summary-level information in the data integration literature as well as the role of rank scores in the quantile regression analysis.

Zhang et al. (2020) also proposed an empirical likelihood approach for data integration, using the summary-level data to make constraints on moments on the validation sample. However, they focus on the regression analyses and the empirical likelihood approach cannot be directly applied here in the causal inference framework. Here we treat $\hat{q}_{t,p}^{\text{Conf}}$ and $\hat{\eta}_t^{\text{Conf}}$ as summary-level data obtained from the entire sample without the detailed confounders S. Equation (3.6) is just obtained from the moment conditions based on the summary-level information. When the entire sample size N is extremely large, the uncertainty in $\hat{q}_{t,p}^{\text{Conf}}$ and $\hat{\eta}_t^{\text{Conf}}$ can be ignored, therefore, we can simply treat them as the true parameters $q_{t,p}^{\text{Conf},*}$ and η_t^*. The right side of (3.6) is simply replacing the expectation in the left side of (2.5) with the empirical measure based on the validation sample.

We can also interpret $\hat{C}_t$ as doubly robust rank scores. Substituting ϕ_t

in (3.6) with its complete expression in (2.4), we obtain that

$$\hat{C}_t = 1/n \sum_{i=1}^{n} \left[\hat{z}_{t,i} \left\{ I(Y_i \leqslant \hat{q}_{t,p}^{\text{Conf}}) - \tilde{G}_t(\hat{q}_{t,p}^{\text{Conf}} \mid X_i; \hat{\theta}_t^{\text{Conf}}) \right\} + \tilde{G}_t(\hat{q}_{t,p}^{\text{Conf}} \mid X_i; \hat{\theta}_t^{\text{Conf}}) - p \right] \tag{3.10}$$

Here $I(Y_i \leqslant \hat{q}_{t,p}^{\text{Conf}}) - p$ serves as the rank score in the quantile regression literature (Koenker, 2005) up to a constant involving the marginal density of Y_i. Consequently, it preserves the main information contained in $q_{t,p}^{\text{Conf}}$. Here in (3.10), $\hat{C}_t$ can be interpreted as a linear combination of doubly robust rank scores where a pseudo propensity score model and an outcome model are included to improve its robustness. Giessing and Wang (2021) use the information in rank scores to debias the conditional quantile treatment effects, while here we use it to integrate information from two datasets.

Our method is also closely related to that in Yang and Ding (2020). They consider the same data configuration as our paper but focus on estimating the ATE, $\tau = E\{Y(1) - Y(0)\}$. The common ground between our method and theirs is that we both project the initial estimators on a consistent estimator for zero. However, a significant difference is that we connect the two datasets through estimating functions rather than simple differences to produce a consistent estimator for zero, which makes it easily adapted to the estimation of QTE. Specifically, suppose that τ is identified by $E\left\{\varphi(U; \tau, \gamma)\right\} = 0$ with a nuisance γ. Yang and Ding (2020) proposed to project the initial estimator on the difference between two error-prone ATE estimators, $\hat{\tau}_{\text{ep}}^{\mathcal{V}}$ and $\hat{\tau}_{\text{ep}}^{\mathcal{O}}$, which are obtained by solving the empirical version of $E\{\varphi(U;\tau,\gamma)\} = 0$ based on the validation data and the entire data separately. Extended to our QTE case, it is equivalent to project on the difference $\hat{C}_{\text{ep}}$ between $\hat{\Delta}_p^{\mathcal{V},\text{Conf}} = \hat{q}_{1,p}^{\mathcal{V},\text{Conf}} - \hat{q}_{0,p}^{\mathcal{V},\text{Conf}}$ and $\hat{\Delta}_p^{\mathcal{O},\text{Conf}} = \hat{q}_{1,p}^{\mathcal{O},\text{Conf}} - \hat{q}_{0,p}^{\mathcal{O},\text{Conf}}$, obtained by solving the empirical version of (2.5) based on two samples separately. Suppose the estimators satisfy

$$n^{1/2}\begin{pmatrix} \hat{\Delta}_p^{\mathcal{V}} - \Delta_p \\ \hat{C}_{\text{ep}} \end{pmatrix} \longrightarrow \mathcal{N}\left\{0, \begin{pmatrix} \sigma_{\mathcal{V}}^2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}. \tag{3.11}$$

We can then obtain a difference-based estimator $\hat{\Delta}_p^{\text{diff}} = \hat{\Delta}_p^{\mathcal{V}} - \hat{\Gamma}^\top \hat{V}^{-1}\hat{C}_{\text{ep}}$ given consistent estimators $\hat{\Gamma}$ and $\hat{V}$.

Unfortunately, due to the essential properties of quantiles, the covariance matrices Γ and V are rather complicated. Heuristically, due to the unmeasured S in the partially-observed data, there are unknown terms related to S appeared simultaneously in Γ and V: the true propensity score $\text{pr}(T = 1|X, S)$, the conditional distribution $F_t(y|X, S)$, as well as the conditional density $f_t(y|X, S)$. Estimation of these terms can be complicated and unstable without additional model assumptions, especially when the validation sample size is relatively small. Although bootstrap method can be used to estimate Γ and V, it can be time-consuming, especially when the entire data size is large, which is exactly the scenario we consider here. See the supplementary materials Section S3.3 for more details and discussions.

However, inspired by the perspective of summary-level data, we can reformulate $\hat{\tau}_{\text{ep}} = \hat{\tau}_{\text{ep}}^{\mathcal{V}} - \hat{\tau}_{\text{ep}}^{\mathcal{O}}$ in Yang and Ding (2020) as $\hat{\tau}_{\text{ep}} = 1/n \sum_{i=1}^{n} \varphi(U_i; \hat{\tau}_{\text{ep}}^{\mathcal{O}}, \hat{\gamma}^{\mathcal{O}})$ by treating $\hat{\tau}_{\text{ep}}^{\mathcal{O}}$ and $\hat{\gamma}^{\mathcal{O}}$ as summary-level data from the dataset with unmeasure confounders, where $\hat{\gamma}^{\mathcal{O}}$ is an estimator for γ using the entire data. Thus, the estimating-function-based connection can be regarded as a generalization of the difference-based connection in Yang and Ding (2020), which, in the case of QTE estimation, leads to $\hat{C}_t$ as in (3.6) with an asymptotical linear representation in (3.7). Combined with the linear form of the initial estimators in (2.3), the covariance matrix can be easily estimated. As we can see from the simulation results in the supplementary materials Section S4.3, besides less computation time, our proposed FQTEs also enjoy better performance than their difference-based estimators.

4. Theoretical Guarantees

4.1 Asymptotic Linear Representation

For the theoretical analysis, let us introduce some additional notations. We use $\xi^{\otimes 2} = \xi\xi^{\top}$ for a vector or matrix ξ. For the outcome model, let $L_t\left(Y, T, X, S; \theta_t\right)$ be the estimating function for θ_t, for $t = 0, 1$. For the propensity score model, let $h(T, X, S; \alpha)$ be the estimating function for α. Moreover, let $\Sigma_\alpha = E\left\{h^{\otimes 2}(T, X, S; \alpha)\right\}$ be the Fisher information matrix for α in the propensity score model.

Simply denote $e_i^* = e\left(X_i, S_i; \alpha^*\right), \dot{e}_i^* = \partial e\left(X_i, S_i; \alpha^*\right)/\partial \alpha^{\mathrm{T}}$, $h_i^* = h\left(T_i, X_i, S_i; \alpha^*\right)$, $\dot{h}_i^* = \partial h\left(T_i, X_i, S_i; \alpha^*\right)/\partial \alpha^{\mathrm{T}}$, $G_{ti}^* = G_t\left(q_{t,p} \mid X_i, S_i; \theta_t^*\right), \dot{G}_{ti}^* = \partial G_t\left(q_{t,p} \mid X_i, S_i; \theta_t^*\right)/\partial \theta_t^{\mathrm{T}}$, $L_{ti}^* = L_t\left(Y_i, T_i, X_i, S_i; \theta_t^*\right)$, and $\dot{L}_{ti}^* = \partial L_t\left(Y_i, T_i, X_i, S_i; \theta_t^*\right)/\partial \theta_t^{\mathrm{T}}$ for $t = 0, 1$. Besides, we denote the density of $Y(t)$ as $f_t(y)$, for $t = 0, 1$. Now we establish the consistency and asymptotic normality of the DR QTE estimators.

Theorem 1. *Under Assumptions 3 or 4, 5, and Condition S3.1 in the supplementary materials Section S3.1, the DR quantile estimators $\hat{q}_{1,p}^{\mathcal{V}}$ and $\hat{q}_{0,p}^{\mathcal{V}}$ obtained by solving (2.2) are asymptotically normal, and (2.3) holds with the influence function*

$$
\begin{aligned}
\psi_1(O_i; q_{1,p}, \eta_1^*) = & -1/f_1(q_{1,p}) \left[T_i\left\{I(Y_i \le q_{1,p}) - G_{1i,p}^*\right\}/e_i^* + G_{1i,p}^* - p\right. \\
& \left. -E\left\{\left(1 - T_i/e_i^*\right)\dot{G}_{1i,p}^*\right\}(E\dot{L}_{1i}^*)^{-1}L_{1i}^* - H_1\Sigma_\alpha^{-1}h_i^*\right], \\
\psi_0(O_i; q_{0,p}, \eta_0^*) = & -1/f_0(q_{0,p}) \left[(1 - T_i)\left\{I(Y_i \le q_{t,p}) - G_{0i,p}^*\right\}/(1 - e_i^*) + G_{0i,p}^* - p\right. \\
& \left. -E\left\{\left(1 - (1 - T_i)/(1 - e_i^*)\right)\dot{G}_{0i,p}^*\right\}(E\dot{L}_{0i}^*)^{-1}L_{0i}^* - H_0\Sigma_\alpha^{-1}h_i^*\right],
\end{aligned}
\tag{4.12}
$$

respectively, where

$$H_1 = E\left[T_i\left\{I(Y_i \le q_{1,p}) - G^*_{1i,p}\right\}\dot{e}^*_i/e^{*2}_i\right],$$

$$H_0 = -E\left[(1-T_i)\left\{I(Y_i \le q_{0,p}) - G^*_{0i,p}\right\}\dot{e}^*_i/(1-e^*_i)^2\right].$$

Consequently, the DR QTE estimator $\hat{\Delta}_p$ is asymptotic linear with influence function $\psi_1(O_i; q_{1,p}, \eta^*_1) - \psi_0(O_i; q_{1,p}, \eta^*_1)$.

Similarly, we establish the asymptotic linear representations of $\hat{C}_t$.

Lemma 1. *Under Assumption 5 and Condition S3.2 in the supplementary materials Section S3.1, we have*

$$\hat{C}_t = 1/n\sum_{i=1}^{n}\phi_t(U_i; q^{\mathrm{Conf},*}_{t,p}, \eta^{\mathrm{Conf},*}_t) - 1/N\sum_{i=1}^{N}\phi_t(U_i; q^{\mathrm{Conf},*}_{t,p}, \eta^{\mathrm{Conf},*}_t) + o_p(1/n^{1/2}),$$

for $t = 0, 1$. That is, the asymptotic linear representation in (3.7) holds.

Remark 3. As we mentioned before, here we allow both $\tilde{G}(X; \theta^{\mathrm{Conf}}_t)$ and $\tilde{e}(X; \alpha^{\mathrm{Conf}})$ to be misspecified for $F_t(y \mid X)$ and $e(X)$ respectively, this is analogous to that in (Yang and Ding, 2020).

4.2 Efficiency Gains

The following theorem shows that our FQTE can improve the efficiency of the initial QTE estimators.

Theorem 2 (Asymptotic Normality)**.** *Under the assumptions in Theorem 1 and Lemma 1, given consistent estimators* $(\hat{\varrho}, \hat{\Sigma}_{\text{ep}})$ *for* $(\varrho, \Sigma_{\text{ep}})$*,* $\hat{\Delta}_p$ *is consistent and*

$$n^{1/2}(\hat{\Delta}_p - \Delta_p) \longrightarrow \mathcal{N}\left(0, \sigma^2\right)$$

in distribution as $n \to \infty$*, where* $\sigma^2 = \sigma^2_{\mathcal{V}} - \varrho^\top \Sigma_{\text{ep}}^{-1} \varrho$*.*

Note that Σ_{ep} is positive definite, which means the FQTE have efficiency gains of $\varrho^\top \Sigma_{\text{ep}}^{-1} \varrho$ compared to the initial estimator, given nonzero ϱ. Theorem 2 also implies that the efficiency gains increase with the covariance between the estimators for zero and the initial QTE estimators based solely on the validation dataset. The higher the covariance ϱ is, the more efficiency gains we will obtain.

4.3 Variance Estimation

Now we discuss how to obtain consistent estimators for the asymptotic covariance matrix in (3.8). The asymptotic linear representations provide us this a convenient way to construct covariance estimators. Consistent estimators for $f_t(q_{t,p})$ are provided in the supplementary materials Section S1. Replacing the unknown terms $(\eta_t, q_{t,p}, f_t(\cdot))$ in $\phi_{t,i}$ and $(\eta_t^{\text{conf}}, q_{t,p}^{\text{conf}})$ in $\psi_{t,i}$ by their correponding estimators $(\hat{\eta}_t, \hat{q}_{t,p}^{\mathcal{V}}, \hat{f}_t(\cdot))$ and $(\hat{\eta}_t^{\text{Conf}}, \hat{q}_{t,p}^{\text{Conf}})$, using empirical measure in place of $E(\cdot)$, we can obtain estimators of $\phi_{t,i}$ and

$\psi_{t,i}$ respectively, denoted by $\hat{\psi}_{t,i}$ and $\hat{\phi}_{t,i}$. Based on Proposition 1, we can estimate $(\varrho, \Sigma_{\text{ep}}, \sigma^2_{\mathcal{V}})$ by

$$\hat{\varrho} = (1-\nu_n)\, 1/n \sum_{i=1}^{n} (\hat{\psi}_{1,i} - \hat{\psi}_{0,i})(\hat{\phi}_{1,i}^{\top}, \hat{\phi}_{0,i}^{\top})^{\top}, \quad \hat{\Sigma}_t = (1-\nu_n)\, 1/N \sum_{i=1}^{N} \hat{\phi}_{t,i}\hat{\phi}_{t,i}^{\top}$$

$$\hat{\Sigma}_{01} = (1-\nu_n)\, 1/N \sum_{i=1}^{N} \hat{\phi}_{0,i}\hat{\phi}_{1,i}^{\top},, \quad \hat{\sigma}^2_{\mathcal{V}} = 1/n \sum_{i=1}^{n} \left(\hat{\psi}_{1,i} - \hat{\psi}_{0,i}\right)^2. \tag{4.13}$$

Based on the consistent estimators for the covariance matrix, we can obtain corresponding variance estimators for our calibrated estimators as

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\mathcal{V}} - \hat{\varrho}^{\top}\hat{\Sigma}_{\text{ep}}^{-1}\hat{\varrho}. \tag{4.14}$$

Given a consistent density estimator, we can establish the consistency of the covariance and variance estimators.

Theorem 3 (Consistent Variance Estimators)**.** *Under Assumptions 3 or 4, and regular conditions in the supplementary materials Section S3.1, given a consistent estimator $\hat{f}_t(y)$ for $t = 0, 1$, the covariance estimators $(\hat{\varrho}, \hat{\Sigma}_{\text{ep}}, \hat{\sigma}^2_{\mathcal{V}})$ in (4.13) are consistent for $(\varrho, \Sigma_{\text{ep}}, \sigma^2_{\mathcal{V}})$. Consequently, the variance estimator $\hat{\sigma}^2$ in (4.14) are consistent for σ^2.*

Remark 4. All the results in Section 3 and 4 are based on the MCAR assumption (Assumption 5). Under a relaxed MAR assumption, where the validation data is no longer a random sample from the entire data, slight modifications to the construction of our FQTE are needed. Analogous results including asymptotic linear representations as well as efficiency gains under the MAR assumption are provided in the supplementary materials Section S2. Technique proofs of all the theorems and lemmas above are provided in the supplementary materials Section S3.

5. Simulation

In this section, we evaluate the empirical performance of our proposed FQTE under the MCAR assumption. Simulation results based on the relaxed MAR assumption as well as a comparison between our FQTE and the direct extension of the difference-based calibration method in (Yang and Ding, 2020) are provided in the supplementary materials Section S4. All the pre-treatment covariates are standardized to have mean 1 and variance 1. We consider a case with 4 confounders, only one commonly observed for the entire data and the rest three observed only for a subset of units. We first generate W_{ki} from the uniform distribution $\mathrm{Unif}(1-\sqrt{3}, 1+\sqrt{3})$, $k = 1, 2, 3$. Let $X_{1i} = W_{1i}, S_{1i} = \exp(W_{2i}/2), S_{2i} = \log(W_{3i}+1), S_{3i} = \sin(3*(W_{1i}))$. The propensity model is set as $\mathrm{logit}\{\mathrm{pr}(T_i = 1 \mid X_i, S_i)\} = 0.25X_{1i} - 0.25S_{1i} + 0.25S_{2i} - 0.25S_{3i}$. Finally set $Y_i(1) = 0.5X_{1i} - 0.5S_{1i} + 0.5S_{2i} - 0.5S_{3i} + \epsilon_i(1)$,$Y_i(0) = 0.5X_{1i} - 0.5S_{1i} + 0.5S_{2i} - 0.5S_{3i} + \epsilon_i(0)$, where

$\epsilon_i(1) \sim \mathcal{N}(0, 2^2)$, $\epsilon_i(0) \sim \mathcal{N}(0, 1)$ and they are independent.

We estimate the QTE at the 0.5th and 0.75th levels. We compare our FQTE $\hat{\Delta}$ using the entire dataset with the initial DR estimators $\hat{\Delta}_p^{\mathcal{V}}$ based solely on the validation data. For each fused estimator, we consider three candidate calibrating quantiles, which are $p_{\text{cal},1} = p$, $p_{\text{cal},2} = (0.5, 0.75)$ and $p_{\text{cal},3} = (0.25, 0.5, 0.75)$ for $p = 0.5, 0.75$ respectively.

Wald-type 95% confidence intervals are constructed based on the variance estimates to compare the empirical coverage rates. Sample sizes (N, n) vary from $(2000, 500)$,$(2000, 1000)$ to $(5000, 1000)$ to show effects of increasing N and n respectively. The results in each scenario are based on 2000 replications. We use "method_v" to stand for the initial QTE estimator based on the validation data only, and "method_c_i" $(i = 1, 2, 3)$ for our fused estimators based on the pseudo quantile $p_{\text{cal},i}$. We also use "dr11", "dr10", "dr01", "dr00" to represent the DR estimators with both OR and PS models correctly specified, only the OR model misspecified, only the PS model misspecified, and both misspecified respectively.

We find that a one-dimensional calibrating quantile is adequate for efficiency improvement, so here we only display the results with a single calibrating quantile. See the supplementary materials Section S4 for additional simulation details and results.

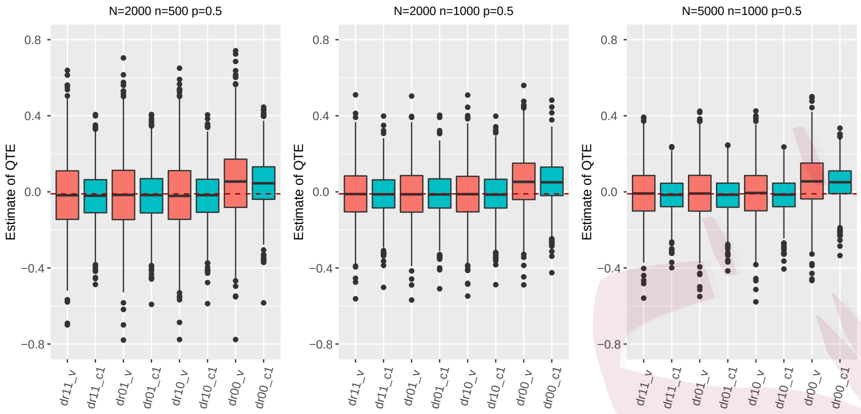

Figure 1: Point estimates of the DR estimators for $\Delta_{0.5}$. Here "dr_v" represents the initial DR QTE estimator and "dr_c_1" represents the DR FQTE using a single quantile.

Figure 1 displays the boxplots of 8 FQTEs for $\Delta_{0.5}$, with the red boxes representing the initial estimators and green boxes representing the FQTEs. The red dash line in the plots represents the true value. Table 1 displays the absolute average bias (BIAS), mean squared error (MSE), standard error (SE) calculated by the variance estimates, and the coverage rate (CR) of our Wald-type 95% confidence intervals for $\Delta_{0.5}$. Results for estimating $\Delta_{0.75}$ are similar and put in the supplementary materials. Our FQTEs enjoy a large gain of efficiency compared to the initial estimators. Specifically, when $N = 2000$ and $n = 500$, the MSEs are reduced by half and the SEs are reduced by a third after data fusion. The BIASes increase slightly after data

Table 1: Simulation Results for estimating $\Delta_{0.5}$. Here "dr_v" represents the initial DR QTE estimator and "dr_c_1" represents the DR FQTE using a single quantile.

$\Delta_{0.5}$	$N = 2000, n = 500$				$N = 2000, n = 1000$				$N = 5000, n = 1000$			
Method	BIAS	MSE	SE	CR	BIAS	MSE	SE	CR	BIAS	MSE	SE	CR
dr11_v	0.0091	0.0358	0.2017	0.9565	0.0001	0.0185	0.1406	0.9610	0.0028	0.0188	0.1407	0.9535
dr11_c1	0.0131	0.0172	0.1418	0.9655	0.0021	0.0119	0.1139	0.9560	0.0063	0.0083	0.0943	0.9605
dr01_v	0.0073	0.0373	0.2042	0.9590	0.0005	0.0191	0.1424	0.9580	0.0024	0.0193	0.1424	0.9560
dr01_c1	0.0112	0.0186	0.1453	0.9640	0.0016	0.0126	0.1160	0.9600	0.0067	0.0089	0.0969	0.9560
dr10_v	0.0080	0.0361	0.2022	0.9605	0.0005	0.0185	0.1410	0.9610	0.0031	0.0189	0.1411	0.9520
dr10_c1	0.0119	0.0173	0.1425	0.9665	0.0027	0.0120	0.1144	0.9570	0.0060	0.0083	0.0948	0.9670
dr00_v	0.0566	0.0401	0.2045	0.9540	0.0657	0.0235	0.1425	0.9325	0.0678	0.0238	0.1426	0.9300
dr00_c1	0.0537	0.0196	0.1399	0.9500	0.0637	0.0163	0.1137	0.9160	0.0591	0.0114	0.0923	0.9015

fusion, which may be caused by the estimation of variances. However, since the bias term is ignorable compared to the variance term, we still benefit from data fusion since the SEs are largely decreased. The efficiency gains grow with the sample size of the main data N and become implicit when the sample size n of the validation data is comparable to N. This shows that an a larger auxiliary dataset will help more with the efficiency improvement while its role becomes unimportant when the validation dataset is already large enough, which is in line with common sense.

For "dr11", "dr10" and "dr01", where at least one model is correctly specified, the coverage rates are all around 95%, which suggests the consistency of our variance estimators. The “dr00” estimators are biased since both models are misspecified, and their coverage rates are lower than 95%.

6. Application

In this section, we apply our data fusion method to evaluate the causal effect of smoking during pregnancy on birth weight (Abrevaya, 2001; Almond et al., 2005; Xie et al., 2020). Based on the Natality Data Set published by National Center for Health Statistics, Almond et al. (2005) showed that births of low-birthweight babies result in both economic costs for society and the children themselves. Meanwhile, they reported a reduction of 203.2 grams in birthweight for smokers versus nonsmokers.

Following Abrevaya (2001) and Almond et al. (2005), we focus on the sample of singleton births and mothers who were either white or black, between the ages of 18 and 45, and the residents in Pennsylvania. We limit the sample to infants born in March, June, September, or December. Analysis of other months yields nearly identical results. The resulting sample size is $N = 29958$. The treatment variable T here is the mother’s smoking status during pregnancy, and the outcome variable Y here is the birthweight of infants (in grams). There are 5558 smokers and 24400 nonsmokers in total. Due to economic costs, researchers may be more interested in the causal effect in the lower quantiles of birthweight. So we consider estimating the 0.5th and 0.25th QTE of smoking during pregnancy on birth weight.

At the same time, large surveys cost a lot of money and time to follow up with the participants and collect some important measures. It may be of great value to cut down the sample size of data needed with full confounders. To illustrate the validity of our data fusion method, we construct the main dataset by including only the basic confounders: mother's marital status, mother's race (either black or white), gender of the infant, mother's age, mother's education and the number of prenatal visits. These five confounders are used as full confounders to evaluate the 0.5th QTE in Xie et al. (2020). However, there are additional key confounders not included in their analyses: alcohol use during pregnancy, the average number of drinks per week, and adequacy of care. We construct the validation dataset by selecting random samples from the whole data including all these eight confounders. The sample size n of the validation dataset varies from $2000, 5000$ to 10000. With estimates based on the whole data with all confounders as a benchmark, we compare our fused estimators using both the validation and main datasets with the initial estimators based solely on the validation dataset. For each fused estimator, we take p_{cal} the same as p for calibration and we propose a normal linear model for the outcome and a lo-

Table 2: Point Estimate (Est), Standard Error (SE) and the Wald-type 95% Confidence Interval. Here "dr_v" represents the initial DR QTE estimator and "dr_c_1" represents the DR FQTE using a single quantile.

		$\Delta_{0.5}$			$\Delta_{0.25}$		
n	Method	Est	SE	95%CI	Est	SE	95%CI
2000	dr_v	-226	44.25	[-312.73,-139.27]	-227	43.66	[-312.58,-141.42]
	dr_c1	-206.95	21.75	[-249.59,-164.33]	-219.84	21.16	[-261.32,-178.39]
5000	dr_v	-189	25.68	[-239.32,-138.68]	-182	28.92	[-238.68,-125.32]
	dr_c1	-189.59	15.15	[-219.22,-159.82]	-215.30	19.45	[-253.81,-177.57]
10000	dr_v	-200	17.90	[-235.09,-164.91]	-199	20.57	[-239.33,-158.67]
	dr_c1	-193.24	11.08	[-214.93,-171.48]	-205.98	13.19	[-231.89,-180.19]
N	dr_v	-198	10.05	[-217.70,-178.30]	-205	11.63	[-227.80,-182.20]

gistic model for the propensity score. We use the random forest to estimate the propensity score in density estimation. The results for QTE estimation are displayed in Table 2.

As we can see from Table 2, the DR estimator for the 0.5th QTE based on the whole sample using all the eight confounders is -198, which indicates a reduction of -198 grams in the median of birthweight for smokers and nonsmokers. Similarly, a larger reduction in the 0.25th quantile of the birthweight, which is -205 grams, is reported. What we want to emphasize here is that our FQTE greatly improves the efficiency of the initial estimators. Therefore, the length of confidence intervals after data fusion is reduced significantly. When the validation sample size n is 10000, which is approximately a third of the whole sample size N, the standard error after data fusion is nearly comparable to that of the estimators based on the whole sample with full confounders. This indicates that we can only collect important information on a representative subsample of the whole data and then our data fusion method can still provide as efficient QTE estimator as that obtained by the whole fully-observed data. Consequently, the cost of money and time in large surveys can be greatly reduced.

7. Discussion

7.1 Generalization of the projection idea

What lies at the heart of our method is the connection of two datasets through estimating functions, which is also a consistent estimator for zero. Once these estimating functions are obtained, projection can be done to improve the efficiency of initial estimators. Inspired by this, consider γ^{Conf} to be an d-dimensional vector of parameters identifiable on the joint distribution of $U = (Y, T, X)$ with

$$E\left\{\varphi(U; \gamma^{\mathrm{Conf}}, \zeta^{\mathrm{Conf},*})\right\} = 0, \tag{7.15}$$

where $\zeta^{\mathrm{Conf},*}$ is an unknown nuisance parameter. Note that (2.5) is a special case of (7.15) by taking $\varphi = \phi_t$, $\zeta^{\mathrm{Conf},*}$ as $q_{t,p}^{\mathrm{Conf}}$ and $\zeta^{\mathrm{Conf},*}$ as $\eta_t^{\mathrm{Conf},*}$. Under Assumption 5, based on the estimates $\hat{\zeta}^{\mathrm{Conf}}$ and $\hat{\gamma}^{\mathrm{Conf}}$ obtained from the main data, we can then connect the two datasets based on estimating functions similar to that in (3.6), formed as $\hat{C} = 1/n\sum_{i=1}^{n}\varphi(U_i;\hat{\gamma}^{\mathrm{Conf}},\hat{\zeta}^{\mathrm{Conf}})$. As implied by Theorem 1, the intuition is to choose φ which results in a $\hat{C}$ with a larger variance and as correlated with the initial estimator as possible. The choice of a well-designed φ to make larger efficiency improvement may be discussed in further work.

7.2 Relaxation of Assumption 3

We may weaken Assumption 3 to the correct specification of the conditional quantile with a small modification of the DR estimation equation (2.2). Consider the quantile regression model $Y_i = g_t(X, S; \theta_t) + \epsilon_{t,i}$, for $T_i = t$ $(t = 0, 1)$, where g_t is known with parameter $\theta_t \in \mathbb{R}^{p_x+p_s}$ and $P(\epsilon_{t,i} < 0 \mid X_i, S_i) = p$. Given an estimate $\hat{\theta}_t$ and the residuals $\hat{\epsilon}_{t,i} = Y_i - g_t(X_i, S_i; \hat{\theta}_t)$, we can replace Equation (2.2) with

$$1/n\sum_{i=1}^{n}\Psi_t^{\mathrm{mod}}(O_i; q, \hat{\eta}_t) = 0,$$

where $\Psi_t^{\mathrm{mod}}(O_i; q, \hat{\eta}_t) = \hat{w}_{t,i}\{I(Y_i \leqslant q) - \frac{1}{n_t}\sum_{i:T_i=t} I(\hat{\epsilon}_{t,i} \leq q)\} + \frac{1}{n_t}\sum_{i:T_i=t} I(\hat{\epsilon}_{t,i} \leq q) - p$ and $n_t = \sum_{i=1}^{n} I(T_i = t)$. The modified quantile estimator is also doubly robust in the sense that it is consistent if either the propensity score model or the conditional quantile $g_t(X, S; \theta_t)$ is correctly specified. Asymptotic linear representations can be also established in a parallel way with a more rigorous proof. Similar ideas can be found in Sued et al. (2020).

7.3 Multiple Auxiliary Datasets

We may also extend our method to incorporate multiple auxiliary datasets as that in Yang and Ding (2020). Specifically, let $U^{(k)} = (Y, T, M^{(k)\top})$ denote the partially-observed information on the k-th auxiliary dataset for $k = 1, \dots, K$, where $M^{(k)} \subsetneqq (X^\top, S^\top)^\top$. Let $\mathcal{O}^{(k)}$ denote the index set of the k-th auxiliary dataset with sample size N_k. For the k-th dataset, we inherit notations from Section 2.3 with X replaced by $M^{(k)}$ and add superscript (k). Following similar procedure as that in Section 2.3, we can obtain confounded DR quantile estimators $\hat{q}_{t,p}^{\text{Conf},(k)}$ by solving the corresponding estimation equation $1/N_k \sum_{i \in \mathcal{O}^{(k)}} \phi_t^{(k)}(U_i^{(k)}; q, \hat{\eta}_t^{\text{Conf},(k)}) = 0$. Then similar to (3.6), we can integrate the information from the k-th auxiliary dataset through $\hat{C}_t^{(k)} = 1/n \sum_{i=1}^n \phi_t^{(k)}(U_i^{(k)}; q_{t,p}^{\text{Conf},(k)}, \hat{\eta}_t^{\text{Conf},(k)})$ for $t = 0, 1$. Asymptotical linear representations for $\{\hat{C}_t^{(k)}\}$ can be established based on $\phi_{t,i}^{(k)}$. Let $\hat{C}^{(k)} = (\hat{C}_1^{(k)\top}, \hat{C}_0^{(k)\top})$ and $\hat{C} = (\hat{C}^{(1)\top}, \dots, \hat{C}^{(K)\top})$. We can also obtain (3.8) with ϱ and Σ_{ep} depending on $\psi_{t,i}$ and $\{\phi_{t,i}^{(k)}\}_{k=1}^K$. Then the FQTE integrating multiple datasets can be similarly obtained by (2.1).

7.4 Sensitivity Analysis

In the real world, the heterogeneity may be intrinsic between the two samples, and hence $\hat{C}_t$ in (3.6) may not converge to 0. Inspired by Yang and Ding (2020), we can introduce a sensitivity parameter δ to quantify the systematic heterogeneity and replace (3.8) with

$$n^{1/2}\begin{pmatrix} \hat{\Delta}_p^{\mathcal{V}} - \Delta_p \\ \hat{C} - \delta \end{pmatrix} \longrightarrow \mathcal{N}\left\{0, \begin{pmatrix} \sigma_{\mathcal{V}}^2 & \varrho^\top \\ \varrho & \Sigma_{\text{ep}} \end{pmatrix}\right\}.$$

The modified estimator becomes $\hat{\Delta}_p^{\text{mod}}(\delta) = \hat{\Delta}_p^{\mathcal{V}} - \hat{\varrho}^\top \hat{\Sigma}_{\text{ep}}^{-1}(\hat{C} - \delta)$. In this way, an investigator is able to assess the impact of the heterogeneity of the two data by varying the values of δ.

Supplementary Materials

The supplementary materials contain extensions to allow for a missing at random mechanism, technical proofs, and additional simulation results.

Acknowledgements

We sincerely thank the editor, associate editor, and two anonymous reviewers for their insightful comments. This work is partially supported by the

National Natural Science Foundation of China grants 12071087.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics 26*, 247–257.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics 120*(3), 1031–1083.

Cai, H., W. Lu, and R. Song (2021). Coda: Calibrated optimal decision making with multiple data sources and limited outcome. *arXiv preprint arXiv:2104.10554*.

Chatterjee, N., Y.-H. Chen, P. Maas, and R. J. Carroll (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association 111*(513), 107–117.

Cheng, D. and T. Cai (2021). Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang (2020). Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*.

Díaz, I. (2017). Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference 190*, 39–51.

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*, 383–397.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Giessing, A. and J. Wang (2021). Debiased inference on heterogeneous quantile treatment effects with regression rank-scores. *arXiv preprint arXiv:2102.01753*.

Han, P., L. Kong, J. Zhao, and X. Zhou (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(2), 305–333.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Kallus, N., A. M. Puli, and U. Shalit (2018). Removing hidden confounding by experimental grounding. *Advances in neural information processing systems 31*.

Koenker, R. (2005). *Quantile Regression.* Econometric Society Monographs. Cambridge University Press.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

Li, S. and A. Luedtke (2021). Efficient estimation under data fusion. *arXiv preprint arXiv:2111.14945*.

Lin, H.-W. and Y.-H. Chen (2014). Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *American journal of epidemiology 180*(3), 308–317.

Rosenman, E. T., A. B. Owen, M. Baiocchi, and H. R. Banack (2022). Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine 41*(1), 65–86.

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

Sued, M., M. Valdora, and V. Yohai (2020). Robust doubly protected estimators for quantiles with missing data. *TEST 29*(3), 819–843.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer New York.

Wang, W., D. Scharfstein, Z. Tan, and E. J. MacKenzie (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 947–969.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1–25.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*(453), 185–193.

Wüthrich, K. (2019). A closed-form estimator for quantile treatment effects with endogeneity.

Journal of Econometrics 210(2), 219–235.

Xie, Y., C. Cotton, and Y. Zhu (2020). Multiply robust estimation of causal quantile treatment effects. *Statistics in Medicine 39*(28), 4238–4251.

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*(531), 1540–1554.

Yang, S., D. Zeng, and X. Wang (2020). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*.

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*(3), 689–703.

Zhang, Z., Z. Chen, J. F. Troendle, and J. Zhang (2012). Causal inference on quantiles with an obstetric application. *Biometrics 68*(3), 697–706.

Department of Statistics and Data Science, Fudan University

E-mail: (yijiaozhang20@fudan.edu.cn)

Department of Statistics and Data Science, Fudan University

E-mail: (zhuzy@fudan.edu.cn)