

# Interpretable Causal Inference for Analyzing Wearable, Sensor, and Distributional Data

Srikar Katta<sup>1</sup>, Harsh Parikh<sup>2</sup>, Cynthia Rudin<sup>1</sup>, and Alexander Volfovsky<sup>1</sup>

<sup>1</sup>Duke University

<sup>2</sup>Johns Hopkins University

December 17, 2023

## Abstract

Many modern causal questions ask how treatments affect complex outcomes that are measured using wearable devices and sensors. Current analysis approaches require summarizing these data into scalar statistics (e.g., the mean), but these summaries can be misleading. For example, disparate distributions can have the same means, variances, and other statistics. Researchers can overcome the loss of information by instead representing the data as distributions. We develop an interpretable method for distributional data analysis that ensures trustworthy and robust decision-making: Analyzing Distributional Data via Matching After Learning to Stretch (ADD MALTS). We (i) provide analytical guarantees of the correctness of our estimation strategy, (ii) demonstrate via simulation that ADD MALTS outperforms other distributional data analysis methods at estimating treatment effects, and (iii) illustrate ADD MALTS' ability to verify whether there is enough cohesion between treatment and control units within subpopulations to trustworthily estimate treatment effects. We demonstrate ADD MALTS' utility by studying the effectiveness of continuous glucose monitors in mitigating diabetes risks.

## 1 Introduction

Diabetes – a disease limiting glucose regulation in the bloodstream – affects millions worldwide. According to the World Health Organization, Diabetes caused 2 million deaths in 2019 and is a leading cause of blindness, kidney failure, and heart attacks (WHO, 2023). Continuous glucose monitors (CGMs), which are wearable devices that automatically track patients' blood glucose concentrations over time, offer a new avenue for diabetes care. CGMs allow researchers and clinicians to screen patients, propose treatments, and manage diets (Matabuena et al., 2021; Janine Freeman et al., 2008; Hall et al., 2018; Lu et al., 2021).

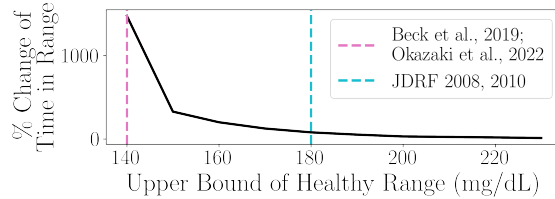


Figure 1: For patients older than 55 years, we measure the effectiveness of CGMs as the percent change of time in healthy range. The plot shows how changing the healthy range’s upper bound (x-axis) affects the treatment effect (y-axis). 70 is the lower bound.

While CGMs show much promise for diabetes care, the standard approaches for summarizing CGM data can lead to very misleading insights. To demonstrate these issues, we reanalyze CGM data from a study conducted by the Juvenile Diabetes Research Foundation (JDRF). JDRF ran a randomized experiment to investigate the effectiveness of CGMs in mitigating the risks of Diabetes using a cohort of 450 patients with type 1 diabetes. Each CGM’s continuous stream of data was summarized by measuring how often a patient’s blood glucose concentration was within a healthy range of 70-180 mg/dL. The treatment effect was then calculated by comparing pre-and-post “time in range” (TIR) between treated and control patients. While JDRF researchers used 70-180 mg/dL as a healthy range, slightly changing the healthy range to 70-140 mg/dL – as used by Okazaki et al. (2022); Beck et al. (2019) – completely changes the results. As shown in Figure 1, for patients older than 55 years old, **using the 70-140 mg/dL range would suggest that CGMs are 1300 percentage points more effective than if the healthy range was 70-180 mg/dL.** This case study highlights how summarizing complex CGM data using scalar statistics can lead to misleading insights, which may be detrimental to patient care.

To overcome the issues of scalar metrics, several researchers have recommended representing data from wearable devices as *distributions* (Matabuena et al., 2021; Ghosal et al., 2023; Ghodrati and Panaretos, 2022). Rather than asking, “how often is a patient’s glucose concentration in a *pre-described* healthy range,” the distributional representation answers the question, “how often is a patient’s glucose concentration at *any particular level for all possible levels.*” Matabuena et al. (2021) demonstrates that the distributional representation of glucose concentrations is much richer than TIR, is clinically useful, and is highly correlated with other clinical biomarkers.

Taking inspiration from the optimal transport literature (Vallender, 1974), Lin et al. (2023) propose estimands and estimators for conducting causal inference with distributional outcomes, enabling us to derive rich insights from CGM data. However, these approaches rely on strong and often untestable assumptions. For example, the positivity assumption requires enough cohesion between treated and control units across subregions of the covariate space. When such assumptions fail, these techniques can yield misleading insights. For proper diabetes care and management, researchers require techniques that can help validate whether the strict assumptions in causal infer-

ence can hold. To this aim, we develop an end-to-end interpretable causal approach for analyzing distributional data: Analyzing Distributional Data via Matching After Learning to Stretch (ADD MALTS).

**Contributions** We prove that ADD MALTS can consistently estimate treatment effects with complex, distributional data. Via simulation, we demonstrate that ADD MALTS can more accurately estimate conditional average treatment effects than competing methods; we also show how ADD MALTS adds trustworthiness in the causal pipeline by validating whether treated and control units are comparable in subregions of the covariate space. Finally, we re-analyze data studying the effectiveness of CGMs in managing health risks in patients with type 1 diabetes, finding important insights about the data and CGMs.

## 2 Background

In this section, we introduce background concepts that are necessary for ADD MALTS. We first discuss the Wasserstein distance, which measures distances between distributions. Next, we discuss how we can “average” distributions (referred to as the barycenter). Finally, we connect the concepts from the Wasserstein space to ideas from traditional causal inference.

**Wasserstein Space** Our work relies on the Wasserstein metric space for measuring distances between distributions (Vallender, 1974). The 2-Wasserstein distance  $W_2(\mu, \nu)$  measures how different cumulative distribution functions (CDFs)  $\mu, \nu$  are from each other by asking how we can transport the mass in  $\mu$  to  $\nu$  in the most cost-effective manner (Panaretos and Zemel, 2019). When distributions are one-dimensional (the focus of our work), the most efficient way of transporting mass between distributions is through their quantiles:  $W_2(\mu, \nu) = \left( \int_0^1 \|\mu^{-1}(q) - \nu^{-1}(q)\|^2 dq \right)^{\frac{1}{2}}$ , where  $\mu^{-1}(q) = \inf\{x \in \mathbb{R} : \mu(x) \geq q\} \forall q \in [0, 1]$  represents the quantile function of  $\mu$ . The quantile function returns the value  $x$  such that the probability of observing a value less than  $x$  is at least as much as  $q$ , the given input probability. Additionally, we can “average” distributions using the Wasserstein distance, referred to as barycenters. Specifically, the Wasserstein barycenter of a set of distributions is the distribution that minimizes the average distance between it and all distributions in the set – the centroid of the set:  $\mathbb{B}[F_Y] \in \arg \min_{\gamma} \mathbb{E}[W_2(F_{Y_i}, \gamma)]$ . With continuous, one-dimensional distributions, the quantile function of the Wasserstein barycenter also has a closed form solution:  $\mathbb{B}[F_Y]^{-1}(q) = \mathbb{E}[F_Y^{-1}(q)]$ . In other words, the quantile function of the average of distributions is the average of quantile functions. We exploit this geometry and represent distributional data via quantiles.

In our setting, we observe  $\mathcal{S}_n$ , a collection of  $n$  independent and identically distributed observations. Each unit  $i$  in  $\mathcal{S}_n$  is assigned to a binary treatment  $T_i \in \{0, 1\}$ ; for notational convenience, let

$\mathcal{S}_n^{(t)}$  represent the set of units whose assigned treatment is  $t$ . We let  $F_{Y_i(1)}$  and  $F_{Y_i(0)}$  represent the treated and control potential outcomes, respectively. We make the standard Stable Unit Treatment Value Assumption (SUTVA); specifically, let  $F_{Y_i} = F_{Y_i(1)}$  if  $t_i = 1$  and  $F_{Y_i} = F_{Y_i(0)}$  if  $t_i = 0$  (Rubin, 2005). Unlike traditional causal inference that assumes the outcomes exist in some Euclidean space, we consider the setting in which the outcomes are continuous distribution functions in the 2-Wasserstein metric space on the closed interval  $\mathcal{I} = [\zeta_{\min}, \zeta_{\max}] \subset \mathbb{R}$ , denoted as  $\mathcal{W}_2(\mathcal{I})$ . For any cumulative distribution function  $F_Y \in \mathcal{W}_2(\mathcal{I})$ ,  $F_Y(s) = 0$  for all  $s \leq \zeta_{\min}$  and  $F_Y(s) = 1$  for all  $s \geq \zeta_{\max}$ . Additionally, let  $F_{\mathbf{X}_i} = [F_{x_{i,1}}, \dots, F_{x_{i,d}}]$  represent a vector of  $d$  *distributional* covariates for unit  $i$  with supports contained within the compact set  $\mathcal{J} \subset \mathbb{R}$ .

*Remark 1.* Because any scalar can be represented as a degenerate distribution, ADD MALTS can handle distribution-on-scalar, scalar-on-distribution, scalar-on-scalar, *and* distribution-on-distribution regression.

Similar to Lin et al. (2023) and Gunsilius (2023), we measure the treatment effect as a contrast between the quantile functions of the potential outcomes. Specifically, we define the individual treatment effect (ITE) as  $\tau_i(q) = F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q)$  for all  $q \in [0, 1]$ . We then define the conditional average treatment effect (CATE) and average treatment effect (ATE) by averaging ITEs: respectively,  $\tau(q|F_{\mathbf{x}}) = \mathbb{E}[\tau_i(q)|F_{\mathbf{X}_i} = F_{\mathbf{x}}]$  and  $\tau(q) = \mathbb{E}[\tau_i(q)]$  for all  $q \in [0, 1]$ , where the expectations are over the observed population.

*Remark 2.* This estimand is different than the quantile treatment effect studied in scalar causal inference (Lin et al., 2023): while quantile treatment effects measure the distribution of differences between potential outcomes, our treatment effect measures the difference between distributional potential outcomes.

We consider the setting where each unit’s treatment assignment and the observed potential outcome may depend on common covariates, referred to as confounders. Under the following assumptions, we can identify (conditional) average treatment effects. First, we assume conditional ignorability, i.e., that the potential outcomes are independent of the assigned treatment given the confounders:  $(F_{Y_i(1)}, F_{Y_i(0)}) \perp\!\!\!\perp T_i \mid F_{\mathbf{X}_i}$ . Additionally, we assume positivity, i.e., that every unit could be in the treated/control group with some chance:  $0 < \mathbb{P}(T_i = 1 \mid F_{\mathbf{X}_i} = F_{\mathbf{x}}) < 1$ . Under these assumptions, we can identify CATEs/ATEs (see Proposition 1 in Section A of the supplement).

## 2.1 Related Literature

Lin et al. (2023) present three strategies for estimating these treatment effects: outcome regression, propensity score weighting, and a doubly robust approach. One outcome regression scheme is to treat the distributional outcome as functional and use functional data analysis tools (Morris, 2015). Similarly, another approach is to predict each quantile of the outcomes using a separate regression (Lin et al., 2023). However, neither of these approaches can guarantee that the predicted distribu-

tional outcome is actually a *distribution*, i.e., integrates to one, with quantile function monotonically increasing. Without these constraints, the imputed counterfactual may not be a distribution.

Other regression approaches combine traditional statistical ideas and take advantage of the linearity of Wasserstein space for univariate distributions via quantile functions. For example, Petersen and Müller (2019); Ghodrati and Panaretos (2022) generalize linear models. Ghosal et al. (2023); Chen et al. (2021); Yang (2020) adapt spline methods. Tang et al. (2023) introduce an expectation-maximization style algorithm. And Qiu et al. (2022) adapt tree algorithms. However, these outcome regression methods are highly sensitive to model misspecification.

Lin et al. (2023) propose an augmented inverse propensity weighting style method that requires only one of the propensity score or outcome regression models to be correctly specified. However, these approaches do not allow for any type of meaningful validation of the important causal assumptions. For example, violations of the positivity assumption significantly reduce the precision of our treatment effect estimates. In order to validate the positivity assumption, researchers often prune observations that have extremal estimated propensity scores (Stuart, 2010; Crump et al., 2009). As we demonstrate in Section 4.2, this strategy is incapable of validating this assumption when the propensity score model is incorrectly specified.

Our approach extends the family of Almost Matching Exactly (AME) methods to the setting of distributional data (Diamond and Sekhon, 2013; Dieng et al., 2019; Parikh et al., 2022; Lanners et al., 2023; Morucci et al., 2023). AME methods learn a distance metric in the covariate space in order to group units that are similar on important covariates; in doing so, we create localized balance, overcoming confounding and enabling us to estimate treatment effects. The conceptual simplicity of these methods makes them easily interpretable and accessible to non-technical audiences. Furthermore, as shown in Parikh et al. (2023), AME methods can also easily integrate qualitative analyses, better aiding decision-making. We extend the family of AME techniques to the setting of distributional data. Our method is highly flexible, end-to-end interpretable, accurate at CATE estimation, and useful in answering important questions using wearable devices, sensors, and other distributional data.

### 3 Methods

**Distance Metric with Distributional Covariates** AME methods learn a distance metric in the covariate space to ensure that matched units are most similar on important features. We first extend the notion of a distance metric to the setting of distributional covariates. Let  $d_{\mathcal{M}}$  represent a distance metric parameterized by the  $d \times d$  diagonal matrix  $\mathcal{M}$ . We measure the distance between unit  $i, j$ 's covariates as

$$d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j}) = \sum_{l=1}^d \mathcal{M}_{l,l} W_2^2(F_{x_{i,l}}, F_{x_{j,l}}).$$

*Remark 3.* Continuous values can be represented as degenerate distributions; and discrete values can be one-hot-encoded and then represented as degenerate distributions. Because the 2-Wasserstein distance between degenerate distributions is the same as the  $\ell_2$  distance between their scalar counterparts, this distance metric can be viewed as a distributional generalization of a weighted Euclidean distance.

**Distance Metric Learning** Traditional nearest neighbor/caliper matching techniques run into the curse of dimensionality where the distance metric can be dominated by less useful covariates when there are many of them (Diamond and Sekhon, 2013). Instead, we *learn* a distance metric and overcome these issues. We first split our data into a training set  $\mathcal{S}_{tr}$  and an estimation set  $\mathcal{S}_{est}$ ; our training and estimation sets are disjoint, so our causal inference remains “honest” and helps lower bias (Rubin, 2005; Athey and Imbens, 2016). On our training set, we evaluate the performance of a proposed distance metric by measuring how well we can predict the observed outcomes; we predict the outcomes for each training unit by averaging the quantile functions of the K-nearest neighbors (KNN) that have the same treatment, where the learned distance metric defines the nearest neighbors.

Under a distance metric  $d_{\mathcal{M}}$ , the KNN of an unit  $i$  in the set of observations with treatment  $t$ ,  $\mathcal{S}^{(t)}$ , is

$$KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}^{(t)}) = \left\{ k : \sum_{j \in \mathcal{S}^{(t)}} \mathbb{1} \left[ d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j}) < d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_k}) \right] < K \right\}. \quad (1)$$

We predict unit  $i$ 's outcome by computing the quantile function of the barycenter of their KNNs' outcomes:

$$\hat{F}_{Y_i}^{-1}(q) = \frac{1}{K} \sum_{j \in KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}^{(t)})} F_{Y_j}^{-1}(q). \quad (2)$$

We then find the optimal distance metric parameters that would yield the best predictions of the observed outcomes using the following objective, a distributional generalization of the mean squared error:

$$\mathcal{M}^*(\mathcal{S}_{tr}) \in \arg \min_{\mathcal{M}} c \|\mathcal{M}\|_{Fr} + \Delta^{(1)}(\mathcal{M}) + \Delta^{(0)}(\mathcal{M}), \text{ where } \Delta^{(t)} = \frac{1}{|\mathcal{S}_{tr}^{(t)}|} \sum_{i \in \mathcal{S}_{tr}^{(t)}} W_2^2(\hat{F}_{Y_i}, F_{Y_i}). \quad (3)$$

Our objective function regularizes the parameters using the Frobenius norm and also considers two treatment-specific loss functions. We evaluate how well we can predict the observed outcomes in the training data by calculating the mean squared Wasserstein distance between the predicted and observed values for the treated units and then the control units.

**CATE Estimation** On the estimation set, we then estimate treatment effects via matching. Specifically, we estimate the quantile functions of the treated and control conditional barycenter for each treatment  $t$  using the set of KNNs:  $\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t]^{-1}(q) = \frac{1}{K} \sum_{j \in \text{KNN}_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}_{est}^{(t)})} F_{Y_j}^{-1}(q)$ . We then estimate the CATE as the difference between the conditional barycenters' quantile functions:

$$\hat{\tau}(q|F_{\mathbf{x}_i}) = \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) - \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q).$$

### 3.1 Theoretical Results

We prove that ADD MALTS consistently estimates conditional barycenters and CATEs by making assumptions analogous to standard ones in the matching literature. Assumption 1 is a Lipschitz continuity-style assumption that states that as the units' covariates become more similar, so too do their conditional barycenters. This is an extension of standard assumptions in the matching literature (Dieng et al., 2019; Parikh et al., 2022; Lanners et al., 2023) to the setting of distributional data and is guaranteed to hold with distributional covariates/outcomes with bounded supports.

**Assumption 1.** *Let  $F_{\mathbf{x}_i}, F_{\mathbf{x}_j} \in \mathcal{W}_2(\mathcal{J})$  and assume  $t_i = t_j$ . If  $d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j}) < \alpha$  for some  $\alpha \in \mathbb{R}$ , then  $W_2(\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t_i], \mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_j}, T = t_j]) < \delta(\alpha)$  for some monotonically increasing, zero-intercept function  $\delta$ .*

The following lemma and theorem rely on this assumption to prove consistency with an intuitive argument. Lemma 1 shows that as we increase the amount of observed data, the radius of each KNN set will decrease. As the radius of each KNN set decreases, the average of the KNN's outcomes will become more similar to that of the query unit's conditional barycenter. This yields the result in Theorem 1: as the estimated conditional barycenters converge to the true conditional barycenters, so too will the estimated CATEs.

**Lemma 1.** *Let Assumption 1 hold. Let  $\varepsilon > 0$  and  $c(\varepsilon, \alpha) = \exp(-K(\varepsilon - 2\delta(\alpha))^2/2(\zeta_{\max} - \zeta_{\min})^2)$ , where  $\alpha$  is the distance of the  $K^{\text{th}}$  nearest neighbor of unit  $i$ . And let  $\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t] \in \arg \min_{\gamma \in \mathcal{W}_2(\mathcal{I})} \frac{1}{K} \sum_{k=1}^K W_2^2(F_{Y_k}, \gamma)$  be the barycenter of the KNN's outcomes. Then,*

$$\mathbb{P}\left(W_2\left(\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t], \mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t]\right) > \varepsilon\right) \leq 2c(\varepsilon, \alpha).$$

**Theorem 1.** *Under the same conditions as Lemma 1,*

$$\mathbb{P}\left(\int_0^1 |\tau(q|F_{\mathbf{x}_i}) - \hat{\tau}(q|F_{\mathbf{x}_i})| dq > \varepsilon\right) \leq 4c\left(\frac{\varepsilon}{2}, \alpha\right).$$

As evident in Theorem 1, as  $\alpha$  decreases and  $K \rightarrow \infty$ , the right hand side will decrease to 0. Because we work on a bounded covariate space,  $\alpha$  is guaranteed to decrease and  $\delta(\alpha)$  will go to 0.

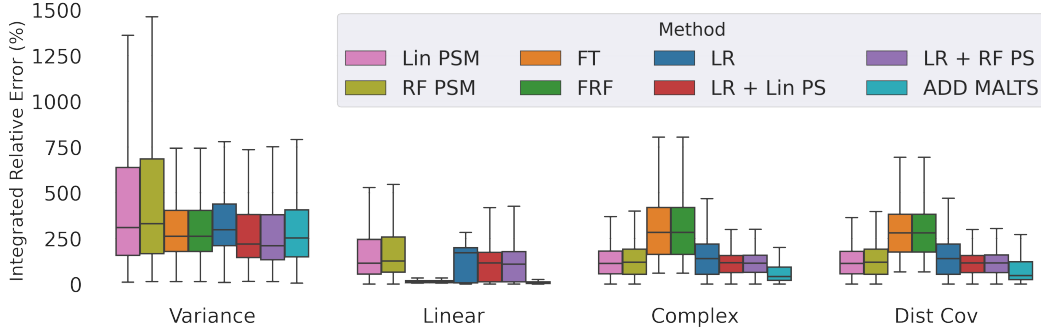


Figure 2: The figure displays the Integrated Relative Error (%) (y-axis) of the different methods we consider for different simulation setups (x-axis). We consider the following baseline methods: **Lin PSM** and **RF PSM** represent propensity score matching fit with linear and random forest models, respectively; **FT** and **FRF** represent decision tree and random forest methods for functional outcomes (Qiu et al., 2022); **LR** represents outcome regression fit at each quantile with a linear regression (Lin et al., 2023); **LR + Lin PS** and **LR + RF PS** represent augmented inverse propensity weighting methods combining the linear outcome regression with linear and random forest propensity score models, respectively.

Therefore, we can consistently estimate conditional barycenters and conditional average treatment effects.

## 4 Simulation Experiments

Our experiments investigate elements essential for causal inference with distributional data: accuracy and trustworthiness. Section 4.1 shows that ADD MALTS estimates CATEs more accurately than baselines and illustrates that ADD MALTS handles scalar *and* distributional covariates. Section 4.2 highlights that ADD MALTS can assess positivity violations.

### 4.1 CATE Estimation

Our first experiment evaluates how well a variety of baselines and ADD MALTS can estimate CATEs. We consider four data generative processes (DGPs). In each DGP, we generate our distributional outcomes as truncated normal distributions (truncated at  $\pm 3$  standard deviations from the mean). In the “Linear,” “Variance,” and “Complex” DGPs, we sample *scalar* covariates while the “Dist Cov” DGP has scalar covariates and one *distributional* covariate. A summary of the distributional covariate is used to generate the distributional outcome in “Dist Cov.” We evaluate each methods ability to estimate the CATE,  $\tau(q|F_{\mathbf{x}_i})$ , using the percent Integrated Relative Error:  $IRE = 100 \times \int_0^1 \left| \frac{\hat{\tau}(q) - \tau(q)}{\tau(q)} \right| dq$ . Section “CATE Estimation Experimental Details” of the supplement expands on our experimental setup.

Figure 2 displays the results of our simulations. The y-axis represents the integrated relative



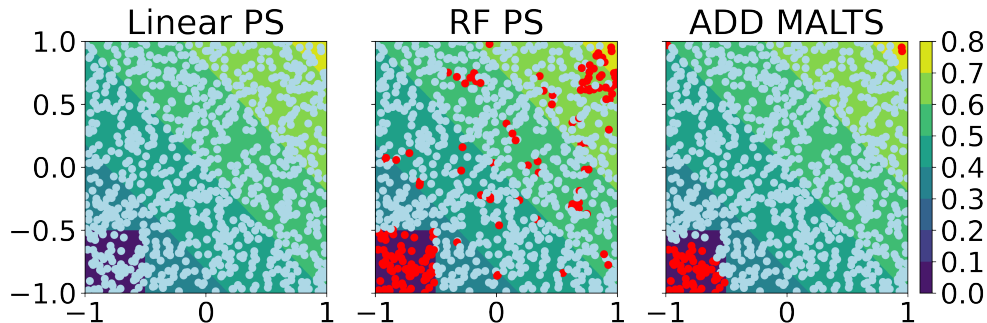


Figure 3: The plot displays which units should be pruned in red according to each method: (from left to right) propensity score estimated with logistic regression using L1 regularization, propensity score estimated with a random forest, and the diameter of matched groups estimated with ADD MALTS. The background displays the true propensity score; the bottom, left corner marks the region of the covariate space with no overlap.

error so smaller values mean better performance. Across the board, ADD MALTS performs at least as well as the baselines (described in the caption). The “Complex” DGP has trigonometric and polynomial relationships between covariates and distributional outcomes. **In this complex setting, ADD MALTS achieves a median IRE of 41.7%, approximately one-third of the error of the next best method (LR + RF PS with median IRE of  $\approx 115\%$ ).**

In the “Dist Cov” DGP, the specific function that summarizes the distributional covariate is the integral over the covariate’s quantile function (see the table in “CATE Estimation Experimental Details” of the supplement). However, in practice, the correct summary function is unknown (e.g., using the mean, median, area under the quantile function). Here, we advantage the baseline methods by letting them use the correct summary value of the distribution (i.e., the integral over its quantile function). On the other hand, ADD MALTS only has access to the raw data drawn from the distribution. **Even in this scenario, ADD MALTS outperforms the baselines.** ADD MALTS again reduces the median IRE from the next best method by one-third, demonstrating that **ADD MALTS can effectively handle both scalar and distributional covariates.** ADD MALTS does not require us to preprocess or summarize distributional covariates, proving that ADD MALTS handles complex data without sacrificing performance.

## 4.2 Positivity Violations

ADD MALTS can also precisely assess positivity violations. While there are several methods for characterizing regions with violations of positivity with scalar data (e.g., Oberst et al., 2020; Crump et al., 2009; Hill et al., 2020), to the best of our knowledge, there are no techniques for distributional data. To benchmark ADD MALTS, we extend methods using estimated propensity scores to our setting. Traditionally, researchers would exclude any units for which the estimated propensity score is not within certain thresholds (e.g., between 0.1 and 0.9) (Stuart, 2010; Crump et al.,

2009; Li et al., 2019). However, these approaches are highly sensitive to model misspecification. We demonstrate that ADD MALTS assesses positivity violations more accurately than these propensity score baselines.

**Using ADD MALTS to Assess Overlap.** First, we match each treated and control unit to their  $K$  nearest neighbors of the opposite treatment status. We calculate each unit’s nearest-neighbor diameter  $D_i$ , the average distance to its nearest neighbors:  $D_i = \frac{1}{K} \sum_{j \in KNN_{d_{\mathcal{M}}}} d_{\mathcal{M}}(F_{\mathbf{x}_i}, \mathcal{S}^{1-t_i})$ . Units with high diameters are far away from units of the opposite treatment and are therefore more likely to be in regions of the covariate space with limited overlap. We flag any units whose diameter is greater than  $D_{\text{upper}} = Q_{D_i}(0.75) + 1.5 \cdot (Q_{D_i}(0.75) - Q_{D_i}(0.25))$ , where  $Q_{D_i}(s)$  is the  $s^{\text{th}}$  quantile of the diameters (this is a common measure of outliers (Suri et al., 2019)).

**Simulation Setup** We simulate data using the following DGP. We have two covariates  $x_{i,0}, x_{i,1} \sim \text{Unif}[-1, 1]$  and the following, piece-wise propensity score model:  $T_i = 0$  if  $x_{i,0} \leq -0.5 \wedge x_{i,1} \leq -0.5$ , else  $T_i \sim \text{Bern}(\text{expit}(-0.5x_{i,0} - 0.5x_{i,1}))$ . In this simulation setup, the true propensity score is linear except for the bottom left corner of the covariate space, where there is no overlap. We train a (parametric)  $\ell_1$ -regularized linear propensity score model (Linear PS) and a (non-parametric) random forest propensity score model (RF PS); any units with propensity scores outside  $[0.1, 0.9]$  were labeled as suffering from positivity violations.

Figure 3 displays the results for one of the iterations of this simulation. The linear propensity score (Linear PS) fails to flag positivity violations. The random forest propensity score (RF PS) correctly characterizes the region of space with positivity violations but at the cost of mischaracterizing regions of overlap as having positivity violations. Dropping observations in regions of the space without positivity violations could adversely affect the precision of our treatment effect estimates. In contrast, ADD MALTS characterizes almost all of the regions of the covariate space correctly. We repeat this experiment 100 times and find that – overall – ADD MALTS accurately classifies 97.6% of units. In comparison, Linear PS and RF PS only classify 93% of units properly. ADD MALTS also enables us to inspect nearest neighbor sets and qualitatively assess whether flagged units suffer from a positivity violation, unlike the propensity score methods. **ADD MALTS precisely flags overlap violations while also being end-to-end interpretable.**

## 5 Real Data Analysis

We use ADD MALTS to reanalyze a clinical trial (Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, 2008) focused on assessing the effectiveness of continuous glucose monitors (CGMs) in mitigating the risk of hyperglycemia (resulting from high glucose levels) or hypoglycemia (linked to low glucose levels) in type 1 diabetes patients. To ensure data validity, we begin by investigating potential violations of the positivity assumption in the CGM trial data.

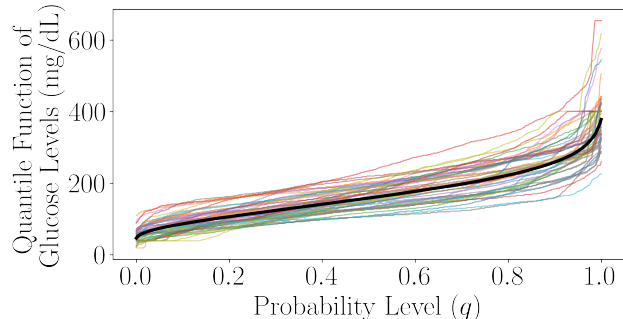


Figure 4: Quantile functions of glucose levels measured at baseline. The thick black line represents the *average* quantile function while the other colors represent the quantile functions for 50 patients.

ADD MALTS detects a lack of positivity for patients who experienced severe hypoglycemia prior to treatment, raising concerns about the generalizability of the trial results to this subgroup. We then assess the heterogeneity of the treatment effects using ADD MALTS. Our findings suggest that, on average, the use of CGMs provides only a marginal benefit in reducing the risk of hyperglycemia or hypoglycemia. However, our subgroup analysis reveals that **CGMs can be beneficial in reducing the risk of extremely high glucose levels for patients aged 55 or older who are effectively managing their diabetes.**

**Data Description** CGMs are wearable devices that monitor blood glucose levels. The Juvenile Diabetes Research Foundation (JDRF) conducted a randomized control trial across 10 clinics and a cohort of 450 patients with type 1 diabetes to assess how helpful CGMs can be in mitigating the risk of extremal glucose concentrations (Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, 2008, 2010). One week prior to randomization, all patients wore modified CGMs where the readings were recorded but not visible to diabetes patients. Patients were then randomly assigned to monitor blood glucose concentrations using CGMs (treatment) or a standard, blood glucose meter (control). The researchers used a stratified randomization scheme to maintain balance based on the clinical center, age group (8 to 14 years, 15 to 24 years, and  $\geq 25$  years), and baseline blood glycated hemoglobin levels ( $HbA1c \leq$  or  $> 8\%$ ). Patients monitored their blood glucose levels using their assigned strategy for 26 weeks; after 26 weeks, all patients wore CGMs with the readings blinded to the control group and visible to the treated group.

**Methods** Previous analyses of these data use the time in range metric to summarize CGM readings (Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, 2008, 2010). As we demonstrate in Figure 1, the time in range metric is highly sensitive to the actual choice of healthy range. To overcome these issues, we re-analyze this data by representing each patient’s CGM data as distributions of glucose concentrations over time (Figure 4 shows the quantile functions of the baseline glucose distributions for 50 patients). We assess overlap and estimate the

Age	HbA1c	Dur.	Col	NHW?	Hypo?	Male	T
43	Low	24.3	True	True	<b>True</b>	True	0
42	High	20.3	True	True	<b>False</b>	True	1
40	Low	24.2	True	True	<b>False</b>	False	1
43	High	20.8	True	True	<b>False</b>	False	1
40	Low	33.1	True	True	<b>False</b>	False	1
41	Low	16.0	True	True	<b>False</b>	True	1

Table 1: The table displays the treated nearest neighbors (bottom five rows) for the query unit (top row).

ATE and CATEs using ADD MALTS.

In our analysis, we include the following control variables. Age describes the unit’s age at randomization (in years). HbA1c describes whether the unit’s glycated blood hemoglobin at baseline was  $\leq 8\%$  (Low) or not (High). “Dur” describes the number of years the patient had diabetes. “Col” denotes whether the patient (or their guardian) graduated from college. “NHW?” is a boolean that is true if the patient is Non-Hispanic White. “Hypo?” denotes whether the patient suffered from an episode of severe hypoglycemia prior to treatment. “Male” denotes whether the patient is male. And Treatment denotes the patient’s treatment assignment. We also control for the distribution of pre-treatment glucose concentrations, as measured by blinded CGMs. We use 40% of the patients to learn ADD MALTS’ distance metric and 60% of the patients to estimate treatment effects.

**Assessing Positivity Violations** We first check for positivity violations in our estimation set. We flagged units in no-overlap regions using the same procedure as in Section 4.2. Table 1 displays the nearest neighbor set for a 43 year old male in the control group who suffered from severe hypoglycemia (i.e., an adverse health outcome due to glucose levels being too low) who was flagged. When we inspect his nearest neighbor set, we see that **there is no similar treated patient who also suffered an episode of severe pre-treatment hypoglycemia**. This unit’s CATE is not very trustworthy, and we would need more data to make such granular insights.

**Estimated Treatment Effects** We first estimate the average treatment effect. As shown in Figure 5, there is little difference between the overall ATE (pink) and the ATE after pruning units suffering from positivity violations (gold). There is a very small, marginal change in glucose concentrations: at each quantile, CGMs only affect glucose levels by between -2 and 1.5 mg/dL, which is miniscule when considering the average person’s glucose readings range between 50 and 350 mg/dL (see Figure 4). **We find that, on average, CGMs do not affect glucose levels.**

ADD MALTS enables us to go beyond ATEs and accurately investigate effects in subpopulations. Using ADD MALTS, we revisit the subpopulation in Figure 1: patients older than 55 years of age. As shown in Figure 5(b), CGMs have marginal effects on glucose concentrations for patients older

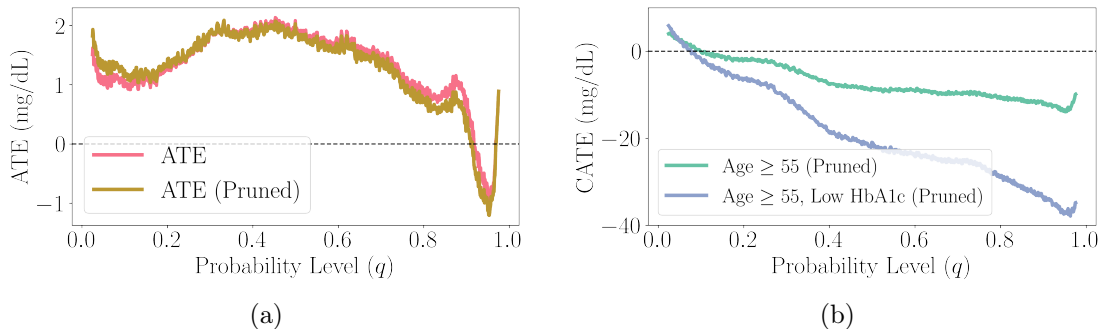


Figure 5: (a) Average treatment effect of glucose-monitoring with CGM on the distribution of glucose concentrations before pruning positivity violations (pink) and after (gold). (b) Conditional average treatment effect of glucose monitoring with CGM on the distribution of glucose concentrations for patients older than 55 years of age (green) and for those with low HbA1c levels at baseline (blue). The x-axes display the probability level and the y-axes display the difference in the related quantiles of the outcomes.

than 55 years of age (green line); extremal glucose levels only change by up to 15 mg/dL. As we see in Figure 4, the average patient’s glucose concentrations range between 50 and 350 mg/dL. A decrease in upper-extremal glucose levels by 15 mg/dL suggests CGMs may be beneficial but not transformative for these older adults’ hyperglycemic risks.

However, CGMs may be very beneficial for patients older than 55 years old who also have low HbA1c levels at baseline; HbA1c is a biomarker that measures the concentration of long-term sugars in the bloodstream. People with lower HbA1c levels tend to have better control of their diabetes needs and lifestyle (i.e., managing diet and exercise). For these patients, upper extremal glucose levels decrease by up to 40 mg/dL, over 10% of the maximum glucose concentration of the average patient (see black line in Figure 4). Because the effect has intensified for patients with low HbA1c levels, this treatment effect estimate suggests that CGMs are most beneficial for patients who actively manage their diabetes needs. On their own, CGMs are not a panacea for diabetes care; however, when coupled with active engagement and self-care, CGMs can help reduce patients’ risk of hyperglycemia.

CGMs tend not to increase *lower* extremal glucose levels, suggesting that CGMs may not be effective in mitigating the risk of hypoglycemia. As suggested by Wolpert (2007), CGMs may cause patients to “over bolus” or overcompensate for rising glucose levels if they observe glucose readings too often. Patients overly concerned about high glucose levels may not manage low glucose concentrations. Understanding that CGMs do not increase low glucose levels suggests CGMs tend not to increase *lower* extremal glucose levels, suggesting that CGMs may not be effective in mitigating the risk of hypoglycemia.

## 6 Conclusion

To enable high-quality and trustworthy causal inference with distributional data, we introduce ADD MALTS. We prove that ADD MALTS can consistently estimate CATEs and validate its performance via simulation. We also show that ADD MALTS effectively handles distributional *and* scalar covariates. ADD MALTS can also precisely flag overlap violations. We use ADD MALTS to study CGMs’ role in glucose level management in type 1 diabetic patients.

**Limitations and Future Directions** While we discuss ADD MALTS’ utility in the context of continuous glucose monitoring data, it can be highly useful for analyzing a variety of other data. For example, distributional representations have been helpful for summarizing images (Oliva et al., 2013; Yang et al., 2020; Zhang et al., 2022) and for summarizing survey data from across geographies (Gunsilius, 2023). However, these complex data may benefit from being represented as *multidimensional* distributions. Future directions should consider extending ADD MALTS to the setting of multidimensional distributional data. Lemma 1 offers an important first step in this direction as our proof of the consistency of the estimated conditional barycenter can easily be extended to multidimensional outcomes. Additionally, uncertainty quantification and variance estimation using distributional outcomes is a difficult and challenging issue that future research should explore. We propose preliminary insights in the supplementary material. Finally, wearable devices are becoming increasingly popular, leading to large datasets with hundreds of thousands of patients (Nazaret et al., 2022). To accommodate such data, future research could benefit from extensions of ADD MALTS that scale to larger datasets without sacrificing accuracy or interpretability.

## References

- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.
- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Beck, R. W., Bergenstal, R. M., Cheng, P., Kollman, C., Carlson, A. L., Johnson, M. L., and Rodbard, D. (2019). The relationships between time in range, hyperglycemia metrics, and HbA1c. *Journal of Diabetes Science and Technology*, 13(4):614–626.
- Chen, Y., Lin, Z., and Müller, H.-G. (2021). Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14.

- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.
- Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.
- Dieng, A., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2019). Interpretable almost-exact matching for causal inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2445–2453. PMLR.
- Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974.
- Ghosal, R., Ghosh, S., Urbanek, J., Schrack, J. A., and Zipunnikov, V. (2023). Shape-constrained estimation in functional regression with bernstein polynomials. *Computational Statistics & Data Analysis*, 178:107614.
- Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica*, 91(3):1105–1117.
- Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., and Snyder, M. (2018). Glucotypes reveal new patterns of glucose dysregulation. *PLoS biology*, 16(7):e2005143.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.
- Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.
- Janine Freeman, R., LD, C., and Lynne Lyons, M. (2008). The use of continuous glucose monitoring to evaluate the glycemic response to food. *Diabetes spectrum*, 21(2):134.
- Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, J. (2008). Continuous glucose monitoring and intensive treatment of type 1 diabetes. *New England Journal of Medicine*, 359(14):1464–1476.
- Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, J. (2010). Effectiveness of continuous glucose monitoring in a clinical care environment: evidence from the juvenile diabetes research foundation continuous glucose monitoring (JDRF-CGM) trial. *Diabetes Care*, 33(1):17–22.

- Lanners, Q., Parikh, H., Volfovsky, A., Rudin, C., and Page, D. (2023). From feature importance to distance metric: An almost exact matching approach for causal inference. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Li, F., Thomas, L. E., and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1):250–257.
- Lin, Z., Kong, D., and Wang, L. (2023). Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):378–398.
- Lu, J., Wang, C., Shen, Y., Chen, L., Zhang, L., Cai, J., Lu, W., Zhu, W., Hu, G., Xia, T., et al. (2021). Time in range in relation to all-cause and cardiovascular mortality in patients with type 2 diabetes: a prospective cohort study. *Diabetes Care*, 44(2):549–555.
- Matabuena, M., Petersen, A., Vidal, J. C., and Gude, F. (2021). Glucodensities: a new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464.
- Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.
- Morucci, M., Rudin, C., and Volfovsky, A. (2023). Matched machine learning: A generalized framework for treatment effect inference with learned metrics. *arXiv preprint arXiv:2304.01316*.
- Nazaret, A., Tonekaboni, S., Darnell, G., Ren, S., Sapiro, G., and Miller, A. (2022). Modeling heart rate response to exercise with wearables data. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*.
- Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., and Varshney, K. (2020). Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR.
- Okazaki, T., Inoue, A., Taira, T., Nakagawa, S., Kawakita, K., and Kuroda, Y. (2022). Association between time in range of relative normoglycemia and in-hospital mortality in critically ill patients: a single-center retrospective study. *Scientific Reports*, 12(1):11864.
- Oliva, J., Póczos, B., and Schneider, J. (2013). Distribution to distribution regression. In *International Conference on Machine Learning*, pages 1049–1057. PMLR.
- Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Applications*, 6:405–431.
- pandas development team, T. (2020). pandas-dev/pandas: Pandas.



- Parikh, H., Hoffman, K., Sun, H., Zafar, S. F., Ge, W., Jing, J., Liu, L., Sun, J., Struck, A., Volfovsky, A., et al. (2023). Effects of epileptiform activity on discharge outcome in critically ill patients in the usa: a retrospective cross-sectional study. *The Lancet Digital Health*.
- Parikh, H., Rudin, C., and Volfovsky, A. (2022). Malts: Matching after learning to stretch. *The Journal of Machine Learning Research*, 23(1):10952–10993.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*.
- Qiu, R., Yu, Z., and Zhu, R. (2022). Random forests weighted local Fréchet regression with theoretical guarantee. *arXiv preprint arXiv:2202.04912*.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.
- Suri, N. M. R., Murty, M. N., and Athithan, G. (2019). *Outlier detection: techniques and applications*. Springer.
- Tang, C., Lenssen, N., Wei, Y., and Zheng, T. (2023). Wasserstein distributional learning via majorization-minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 10703–10731. PMLR.
- Vallender, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.
- WHO, W. H. O. (2023). Diabetes. <https://www.who.int/news-room/fact-sheets/detail/diabetes>.

Wolpert, H. A. (2007). Use of continuous glucose monitoring in the detection and prevention of hypoglycemia. *J Diabetes Sci Technol*, 1:146–150.

Yang, H. (2020). Random distributional response model based on spline method. *Journal of Statistical Planning and Inference*, 207:27–44.

Yang, H., Baladandayuthapani, V., Rao, A. U., and Morris, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *Journal of the American Statistical Association*, 115(529):90–106.

Zhang, Z., Wang, X., Kong, L., and Zhu, H. (2022). High-dimensional spatial quantile function-on-scalar regression. *Journal of the American Statistical Association*, 117(539):1563–1578.

## Appendix A Identification of the ATE

We first prove that we can identify average treatment effects. This is the same theorem statement as in Lin et al. (2023) and the proof follows the same arguments. This is here for completeness.

**Proposition 1** (Lin et al. (2023)). *Under SUTVA, conditional ignorability, and positivity, we can identify average treatment effects.*

*Proof.* We aim to identify the average treatment effect,

$$\tau(q) = \mathbb{E} \left[ F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q) \right],$$

where the expectation is over the population we sample data from.

Let  $q \in [0, 1]$ . By the law of iterated expectations,

$$\begin{aligned} \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q)] &= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}] \right\} \\ &= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}] - \mathbb{E}_i[F_{Y_i(0)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}] \right\}, \end{aligned}$$

by the linearity of expectations. Recall that by conditional ignorability, we know that for all  $q \in [0, 1]$ ,  $\mathbb{E}_i[F_{Y_i(T_i)}^{-1}(q) | F_{\mathbf{x}_i}] = \mathbb{E}_i[F_{Y_i(T_i)}^{-1}(q) | T_i, F_{\mathbf{x}_i}]$ . Substituting this equality into our equation, we know that

$$\begin{aligned} &\mathbb{E}_i[F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q)] \\ &= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}] - \mathbb{E}_i[F_{Y_i(0)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}] \right\} \\ &= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}, T_i = 1] - \mathbb{E}_i[F_{Y_i(0)}^{-1}(q) | F_{\mathbf{X}_i} = F_{\mathbf{x}}, T_i = 0] \right\}. \end{aligned}$$

Because we have now conditioned on observing a specific treatment for our units, we know by SUTVA that  $F_{Y_i} = F_{Y_i(T_i)}$ . So,

$$\begin{aligned}
& \mathbb{E}_i[F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q)] \\
&= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}, T_i = 1] - \mathbb{E}_i[F_{Y_i}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}, T_i = 0] \right\} \\
&= \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}, T_i = 1] \right\} - \mathbb{E}_{F_{\mathbf{X}}} \left\{ \mathbb{E}_i[F_{Y_i}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}, T_i = 0] \right\} \\
&= \mathbb{E}_i[F_{Y_i}^{-1}(q) | T_i = 1] - \mathbb{E}_i[F_{Y_i}^{-1}(q) | T_i = 0],
\end{aligned} \tag{4}$$

where the last two steps follow by using the linearity of expectations and reversing the iterated expectation.  $\square$

## Appendix B CATE Estimation Consistency Proofs

This section establishes the consistency of ADD MALTS' CATE estimation strategy. We first re-introduce the smoothness assumption used in these proofs:

**Assumption 2.** *Let  $F_{\mathbf{x}_i}, F_{\mathbf{x}_j} \in \mathcal{W}_2(\mathcal{J})$  and assume  $t_i = t_j$ . If  $d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j}) < a$  for some  $a \in \mathbb{R}$ , then  $W_p(\mathbb{B}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t_i], \mathbb{B}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_j}, T = t_j]) < \delta(a)$  for some monotonically increasing, zero-intercept function  $\delta$ .<sup>1</sup>*

### B.1 Proof of Barycenter Consistency (Lemma 1)

**Lemma.** *Let Assumption 2 hold. Additionally, let  $-\infty < \zeta_{\min} \leq \zeta_{\max} < \infty$  represent the lower and upper bounds on the support of the space of outcome distributions,  $\mathcal{W}_2(\mathcal{I})$ . Specifically, for any cumulative distribution function  $F_Y \in \mathcal{W}_2(\mathcal{I})$ ,  $F_Y(s) = 0$  for all  $s \leq \zeta_{\min}$  and  $F_Y(s) = 1$  for all  $s \geq \zeta_{\max}$ . Lastly, let  $a$  be the distance of the  $K^{\text{th}}$  nearest neighbor. Then, for all  $\varepsilon > 0$ ,*

$$\mathbb{P} \left( W_p \left( \hat{\mathbb{B}}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t], \mathbb{B}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t] \right) > \varepsilon \right) \leq 2 \exp \left( \frac{-K(\varepsilon - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right),$$

where  $\hat{\mathbb{B}}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t] = \arg \min_{\gamma \in \mathcal{W}_2(\mathcal{I})} \frac{1}{K} \sum_{k=1}^K W_p^2(F_{Y_k}, \gamma)$  is estimated as the barycenter of unit  $i$ 's KNN's outcomes.

*Proof.* We will prove that the following inequality holds for *any*  $p$ -Wasserstein distance,  $p \geq 1$ . So, it will hold for the 2-Wasserstein distance as well.

$$\mathbb{P} \left( W_p \left( \hat{\mathbb{B}}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t], \mathbb{B}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t] \right) > \varepsilon \right) \leq 2 \exp \left( \frac{-K(\varepsilon - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right)$$

---

<sup>1</sup>Please note that this assumption is slightly different than the assumption in the main paper. In the main paper, the assumption requires that this smoothness condition needs to hold only with the 2-Wasserstein distance. This assumption requires the smoothness condition to hold for *any*  $p$ -Wasserstein distance with  $p \geq 1$ .

Please note that we fix the covariates and treatment assignments for all units. The probability distribution of our inequality is with respect to the *observed* outcomes for each unit.

Let  $\mathbb{B}_i = \mathbb{B}[F_Y | F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t]$  represent the conditional barycenter of the outcome at  $F_{\mathbf{x}_i}$  and treatment  $t$ . Let  $KNN_i = KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}_{est}^{(t)})$  represent the  $K$  nearest neighbors to unit  $i$  in the estimation set with assigned treatment  $t$ . We demonstrate that ADD MALTS' estimate –  $\hat{\mathbb{B}}_i = \arg \min_{\gamma \in \mathcal{W}_2(\mathcal{I})} \frac{1}{K} \sum_{k=1}^K W_p^2(F_{Y_k}, \gamma)$  – consistently estimates the true conditional barycenter's quantile function.

Now, let  $\varepsilon > \delta(a)$ . First, notice that by using a series of triangle inequalities, we can rewrite

$$\begin{aligned} & \mathbb{P} \left( W_p \left\{ \mathbb{B}_i, \hat{\mathbb{B}}_i \right\} > \varepsilon \right) \\ &= \mathbb{P} \left( \underbrace{\frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ \mathbb{B}_i, \hat{\mathbb{B}}_i \right\}}_{\text{Mult by 1}} > \varepsilon \right) \\ &\leq \mathbb{P} \left( \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ \mathbb{B}_i, \mathbb{B}_k \right\} + \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ \mathbb{B}_k, \hat{\mathbb{B}}_i \right\} > \varepsilon \right) \\ &\leq \mathbb{P} \left( \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ \mathbb{B}_i, \mathbb{B}_k \right\} + \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ \mathbb{B}_k, F_{Y_k} \right\} + \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ F_{Y_k}, \hat{\mathbb{B}}_i \right\} > \varepsilon \right). \end{aligned}$$

Because  $k \in KNN_i$ , we know that  $d_{\mathcal{M}} \{F_{\mathbf{x}_i}, F_{\mathbf{x}_k}\} < a$ . Then, by Assumption 2, we know that  $W_p \{ \mathbb{B}_i, \mathbb{B}_k \} < \delta(a)$ . So, we can replace

$$\frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_i, \mathbb{B}_k \} \leq \delta(a).$$

Now, realize that  $\hat{\mathbb{B}}_i$  is the centroid of the outcome distributions of the matched group. So, by definition,

$$\begin{aligned} \frac{1}{K} \sum_{k \in KNN_i} W_p \left\{ F_{Y_k}, \hat{\mathbb{B}}_i \right\} &= \min_{\gamma} \frac{1}{K} \sum_{k \in KNN_i} W_p(F_{Y_k}, \gamma) \\ &\leq \frac{1}{K} \sum_{k \in KNN_i} W_p(F_{Y_k}, \mathbb{B}_i) \end{aligned}$$

because  $\mathbb{B}_i$  is either the centroid of the KNN set or is on average further away from any of the distributions in the set. Then, by using the triangle inequality once again, we can decompose this

inequality once more:

$$\begin{aligned}
\frac{1}{K} \sum_{k \in KNN_i} W_p \{F_{Y_k}, \hat{\mathbb{B}}_i\} &= \min_{\gamma} \frac{1}{K} \sum_{k \in KNN_i} W_p \{F_{Y_k}, \gamma\} \\
&\leq \frac{1}{K} \sum_{k \in KNN_i} W_p \{F_{Y_k}, \mathbb{B}_i\} \\
&\leq \frac{1}{K} \sum_{k \in KNN_i} W_p \{F_{Y_k}, \mathbb{B}_k\} + \frac{1}{K} \sum_{k \in KNN_i} W_p \{\mathbb{B}_k, \mathbb{B}_i\} \\
&\leq \delta(a) + \frac{1}{K} \sum_{k \in KNN_i} W_p \{F_{Y_k}, \mathbb{B}_k\},
\end{aligned}$$

where the last step follows from the smoothness assumption. Now, we can combine all of these small inequalities together:

$$\begin{aligned}
&\mathbb{P} \left( W_p \{ \mathbb{B}_i, \hat{\mathbb{B}}_i \} > \varepsilon \right) \\
&\leq \mathbb{P} \left( W_p \{ \mathbb{B}_i, \mathbb{B}_k \} + \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} + \frac{1}{K} \sum_{k \in KNN_i} W_p \{ F_{Y_k}, \hat{\mathbb{B}}_i \} > \varepsilon \right) \\
&\leq \mathbb{P} \left( \delta(a) + \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} + \delta(a) + \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} > \varepsilon \right) \\
&= \mathbb{P} \left( \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} > \frac{\varepsilon - 2\delta(a)}{2} \right).
\end{aligned}$$

Now, recall that  $F_{Y_k} \in \mathcal{W}_2(\mathcal{I})$ , so  $\mathbb{B}_k \in \mathcal{W}_2(\mathcal{I})$ . In this space, the maximum distance occurs when we move the entire distribution's mass (i.e., 1) as far as possible, which is  $|\zeta_{\max} - \zeta_{\min}|$ . So,  $\max_{F_A, F_B} W_p \{F_A, F_B\} = \zeta_{\max} - \zeta_{\min}$ . Additionally, because  $W_p$  is a distance metric, the  $p$ -Wasserstein distance between distributions is bounded from below by 0. Because  $W_p \{ \mathbb{B}_k, F_{Y_k} \}$  is bounded and each unit  $k$  is independent, we can then bound this probability using Hoeffding's inequality: for some  $t > 0$ ,

$$\mathbb{P} \left( \left| \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} \right| > t \sqrt{\frac{(\zeta_{\max} - \zeta_{\min})^2}{K}} \right) \leq 2 \exp(-2t^2).$$

Let  $t = \frac{(\varepsilon - 2\delta(a))}{2\sqrt{\frac{(\zeta_{\max} - \zeta_{\min})^2}{K}}}$ . Then,

$$\mathbb{P} \left( \left| \frac{1}{K} \sum_{k \in KNN_i} W_p \{ \mathbb{B}_k, F_{Y_k} \} \right| > t \sqrt{\frac{(\zeta_{\max} - \zeta_{\min})^2}{K}} \right) \leq 2 \exp(-2t^2)$$

$$\begin{aligned}
&\leq 2 \exp \left( -2 \left[ \frac{\varepsilon - 2\delta(a)}{2\sqrt{\frac{(\zeta_{\max} - \zeta_{\min})}{K}}} \right]^2 \right) \\
&= 2 \exp \left( \frac{-K(\varepsilon - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right)
\end{aligned}$$

□

## B.2 Proof of CATE Consistency (Theorem 1)

**Theorem.** *Under the same conditions as Lemma 1,*

$$\mathbb{P} \left( \int_0^1 |\tau(q|F_{\mathbf{X}_i}) - \hat{\tau}(q|F_{\mathbf{X}_i})| dq > \varepsilon \right) \leq 4 \exp \left( \frac{-K(\frac{\varepsilon}{2} - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right).$$

*Proof.* We estimate conditional average treatment effects as a contrast between the quantile functions of the treated and control conditional barycenters. Specifically,

$$\hat{\tau}(q) = \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) - \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q).$$

We want to show that for all  $q \in [0, 1]$ ,  $\hat{\tau}(q)$  converges to  $\tau(q)$ :

$$\tau(q) = \mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) - \mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q).$$

Let  $\varepsilon > 0$ . By definitions and rearranging,

$$\begin{aligned}
&\mathbb{P} \left( \left\{ \int_0^1 |\hat{\tau}(q) - \tau(q)| dq \right\} > \varepsilon \right) \\
&= \mathbb{P} \left( \left\{ \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} - \begin{pmatrix} \mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} \right| dq \right\} > \varepsilon \right) \\
&= \mathbb{P} \left( \left\{ \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \end{pmatrix} - \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} \right| dq \right\} > \varepsilon \right).
\end{aligned}$$

By the triangle inequality, we see that

$$\begin{aligned}
&\mathbb{P} \left( \left\{ \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \end{pmatrix} - \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} \right| dq \right\} > \varepsilon \right) \\
&\leq \mathbb{P} \left( \left\{ \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \end{pmatrix} \right| + \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} \right| dq \right\} > \varepsilon \right) \\
&= \mathbb{P} \left( \left\{ \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q) \end{pmatrix} \right| dq + \int_0^1 \left| \begin{pmatrix} \hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \\ -\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q) \end{pmatrix} \right| dq \right\} > \varepsilon \right).
\end{aligned}$$

And by the union bound, we see that

$$\begin{aligned} & \mathbb{P} \left( \left\{ \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)} \right| dq + \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)} \right| dq \right\} > \varepsilon \right) \\ & \leq \mathbb{P} \left( \left\{ \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)} \right| dq \right\} > \frac{\varepsilon}{2} \right) + \mathbb{P} \left( \left\{ \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)} \right| dq \right\} > \frac{\varepsilon}{2} \right). \end{aligned}$$

Recall that the 1-Wasserstein distance between one-dimensional distributions is a contrast between quantile functions. So,

$$\begin{aligned} & \mathbb{P} \left( \left\{ \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]^{-1}(q)} \right| dq \right\} > \frac{\varepsilon}{2} \right) + \mathbb{P} \left( \left\{ \int_0^1 \left| \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)}{-\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]^{-1}(q)} \right| dq \right\} > \frac{\varepsilon}{2} \right) \\ & = \mathbb{P} \left( W_1 \left( \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]}{\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]} \right) > \frac{\varepsilon}{2} \right) + \mathbb{P} \left( W_1 \left( \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]}{\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]} \right) > \frac{\varepsilon}{2} \right). \end{aligned}$$

From Lemma 1, we know that

$$\mathbb{P} \left( W_p \left( \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t]}{\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = t]} \right) > \frac{\varepsilon}{2} \right) \leq 2 \exp \left( \frac{-K(\frac{\varepsilon}{2} - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})} \right).$$

Therefore,

$$\begin{aligned} & \mathbb{P} \left( W_1 \left( \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]}{\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 1]} \right) > \varepsilon \right) + \mathbb{P} \left( W_1 \left( \frac{\hat{\mathbb{B}}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]}{\mathbb{B}[F_Y|F_{\mathbf{X}} = F_{\mathbf{x}_i}, T = 0]} \right) > \varepsilon \right) \\ & \leq 4 \exp \left( \frac{-K(\frac{\varepsilon}{2} - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right), \end{aligned}$$

where  $a$  is the maximum distance between unit  $i$  and any of its treated or control  $K$  nearest neighbors.

Therefore,

$$\mathbb{P} \left( \left\{ \int_0^1 |\tau(q|F_{\mathbf{x}_i}) - \hat{\tau}(q|F_{\mathbf{x}_i})| dq \right\} > \varepsilon \right) \leq 4 \exp \left( \frac{-K(\frac{\varepsilon}{2} - 2\delta(a))^2}{2(\zeta_{\max} - \zeta_{\min})^2} \right).$$

□

## Appendix C Uncertainty Quantification

In this section, we provide different strategies for quantifying uncertainty around (i) sample average treatment effect on treated (SATT) and (ii) conditional average treatment effect estimates.

We construct point-wise confidence intervals for the average treatment effect. The motivation and theory follow almost directly from Abadie and Imbens (2011), but we discuss these in the

context of our notation and setting with distributional outcomes.

Let  $\mu_t(q | F_{\mathbf{x}}) = \mathbb{E} \left[ F_{Y_i}^{-1}(q) | F_{\mathbf{X}} = F_{\mathbf{x}}, T = t \right]$  and  $\sigma_t^2(q | F_{\mathbf{x}}) = \mathbb{E} \left[ \left( F_{Y_i}^{-1}(q) - \mu_t(q | F_{\mathbf{x}}) \right)^2 | F_{\mathbf{X}} = F_{\mathbf{x}}, T = t \right]$ . Assume  $\mu_t(q | F_{\mathbf{x}_i})$  and  $\sigma_t^2(q | F_{\mathbf{x}})$  are smooth: for units  $i, j$  with  $t_i = t_j$ ,  $|\mu_t(q | F_{\mathbf{x}_i}) - \mu_t(q | F_{\mathbf{x}_j})|$  and  $|\sigma_t^2(q | F_{\mathbf{x}_i}) - \sigma_t^2(q | F_{\mathbf{x}_j})|$  strictly monotonically decrease as  $d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j})$  decreases and are 0 when  $d_{\mathcal{M}}(F_{\mathbf{x}_i}, F_{\mathbf{x}_j}) = 0$ . This assumption places smoothness at the quantile-level rather than at the quantile-*function* level. Also, assume that  $\mathbb{E} \left[ (F_{Y_i}^{-1}(q))^4 | F_{\mathbf{X}} = F_{\mathbf{x}_i} \right] \leq C$  for some finite  $C$  and that  $\sigma_t^2(q | F_{\mathbf{x}})$  is bounded away from 0.

Our estimand of interest is the average treatment effect (ATE):

$$\tau(q) = \mathbb{E} \left[ F_{Y_i(0)}^{-1}(q) - F_{Y_i(1)}^{-1}(q) \right].$$

Let  $\hat{\mu}_t(q | F_{\mathbf{x}_i})$  represent a consistent estimator of  $\mu_t(q | F_{\mathbf{x}_i})$ . Also, let  $M_K(i) = \sum_{l=1}^{|\mathcal{S}_{est}|} \mathbb{1} \left[ i \in KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_l}, \mathcal{S}_{est}^{(1-t)}) \right]$  represent the number of times unit  $i$  is in *another* unit's set of  $K$  nearest neighbors of the opposite treatment status. Rewriting our estimate for the ATE:

$$\begin{aligned} \hat{\tau}(q) &= \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} (2T_i - 1) \cdot \left( 1 + \frac{M_K(i)}{K} \right) \cdot F_{Y_i}^{-1}(q) \\ &= \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} \hat{F}_{Y_i(1)}^{-1}(q) - \hat{F}_{Y_i(0)}^{-1}(q), \end{aligned}$$

where

$$\hat{F}_{Y_i(t)}^{-1}(q) = \begin{cases} F_{Y_i}^{-1}(q) & \text{if } t = T_i \\ \frac{1}{K} \sum_{j \in KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}_{est}^{(1-t)})} F_{Y_j}^{-1}(q) & \text{if } t \neq T_i. \end{cases}$$

However,  $\hat{\tau}(q)$  is biased; Abadie and Imbens (2011) provide a closed form solution for the bias term and an estimator:

$$B(q) = \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} \frac{2T_i - 1}{K} \sum_{j \in KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}_{est}^{1-T_i})} \mu_{1-T_i}(q | F_{\mathbf{x}_i}) - \mu_{1-T_i}(q | F_{\mathbf{x}_j}).$$

We then have an unbiased estimate of the ATE,  $\tau(q) = \mathbb{E}[\hat{\tau}(q) - B(q)]$ . However, we do not know the true conditional mean function  $\mu_t(q | F_{\mathbf{x}_i})$ , so we must estimate them. Assume we have consistent estimators of the conditional means of the outcomes. Then, we can consistently estimate  $B(q)$ :

$$\hat{B}(q) = \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} \frac{2T_i - 1}{K} \sum_{j \in KNN_{d_{\mathcal{M}}}(F_{\mathbf{x}_i}, \mathcal{S}_{est}^{1-T_i})} \hat{\mu}_{1-T_i}(q | F_{\mathbf{x}_i}) - \hat{\mu}_{1-T_i}(q | F_{\mathbf{x}_j})$$



We can then construct our bias-corrected estimator  $\hat{\tau}^{bcm}(q) = \hat{\tau}(q) - \hat{B}(q)$ .

We now construct the variance of our bias corrected estimator. Let  $\{F_{\mathbf{x}_i}, T_i\}_{i \in \mathcal{S}_{est}}$  represent the covariates and treatment statuses observed in the estimation set. And let

$$\begin{aligned} V^E(q) &= \frac{1}{|\mathcal{S}_{est}|} \mathbb{V} [\hat{\tau}(q) \mid \{F_{\mathbf{x}_i}, T_i\}_{i \in \mathcal{S}_{est}}] \\ &= \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{M_K(i)}{K}\right)^2 \sigma_{T_i}^2(q \mid F_{\mathbf{x}_i}), \\ V^{\tau(q|F_{\mathbf{x}_i})}(q) &= \mathbb{E} \left[ (\mu_1(q \mid F_{\mathbf{x}_i}) - \mu_0(q \mid F_{\mathbf{x}_i}) - \tau(q))^2 \right]. \end{aligned}$$

Then,

$$\sqrt{|\mathcal{S}_{est}|} \left( V^E(q) + V^{\tau(q|F_{\mathbf{x}_i})}(q) \right)^{-1/2} \left( \hat{\tau}^{bcm}(q) - \tau(q) \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

By estimating  $V^E(q) + V^{\tau(q|F_{\mathbf{x}_i})}(q)$ , we could construct  $1 - \alpha\%$  confidence intervals. To do this, we must first find a way to estimate  $\sigma_{T_i}^2(q \mid F_{\mathbf{x}_i})$ , the conditional variance of the outcome. Let  $\ell_j(i)$  represent the  $j^{th}$  nearest neighbor of  $i$  with the same treatment status in the estimation set. For some fixed  $J$ , we can estimate  $\sigma_{T_i}^2(q \mid F_{\mathbf{x}_i})$  as

$$\hat{\sigma}_{T_i}^2(q \mid F_{\mathbf{x}_i}) = \frac{J}{J+1} \left( F_{Y_i}^{-1}(q) - \frac{1}{J} \sum_{j=1}^J F_{Y_{\ell_j(i)}}^{-1}(q) \right)^2.$$

Theorem 7 of Abadie and Imbens (2006) then shows that we can estimate valid confidence intervals using the following consistent estimator of  $V(q) = V^E(q) + V^{\tau(q|F_{\mathbf{x}_i})}(q)$ :

$$\begin{aligned} \hat{V}(q) &= \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} \left| \hat{F}_{Y_i(1)}^{-1}(q) - \hat{F}_{Y_i(0)}^{-1}(q) - \hat{\tau}(q) \right|^2 \\ &\quad + \frac{1}{|\mathcal{S}_{est}|} \sum_{i \in \mathcal{S}_{est}} \left[ \left( \frac{M_K(i)}{K} \right)^2 + \left( \frac{2K-1}{K} \right) \left( \frac{M_K(i)}{K} \right) \right] \hat{\sigma}_{T_i}^2(q \mid F_{\mathbf{x}_i}). \end{aligned} \quad (5)$$

## C.1 Simulation Study

We use the data generating processes with scalar covariates in Table 3 to evaluate the coverage of our uncertainty quantification strategy. For each of 100 Monte Carlo iterations, we first calculate the true average treatment effect and use the variance estimator in Equation 5 to construct 95% *pointwise* confidence intervals. We then evaluate the nominal pointwise coverage.

DGP	Coverage
Variance	0.947
Linear	0.963
Complex	0.989

Table 2: The estimated coverage of 95% confidence intervals using the Standard Interval ( $\hat{V}[\hat{\tau}(q)] = \Sigma$ ) and Error-calibrated Interval (from Equation 5).

## Appendix D Real Data Analysis: Exploratory Data Analysis

In this section, we further explore the data from Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group (2010, 2008). Specifically, we first show how we clean the data; and because the trial randomized treatment, we compare ADD MALTS’ estimated ATE to the difference-in-means-estimated ATE to validate that ADD MALTS can also recover an accurate ATE estimate.

**Data Cleaning and Processing** The data analyses were published in two separate studies: Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group (2008) studies patients for whom glycated blood hemoglobin levels (HbA1c) at baseline was greater than 8% and Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group (2010) studies patients for whom glycated blood hemoglobin levels (HbA1c) at baseline was  $\leq 8\%$ . Both studies excluded patients who did not complete the full 26 weeks of randomization; we also excluded patients that did not have CGM readings more than 26 weeks worth of readings after the last recording in their baseline data. We constructed each quantile function using 900 quantiles; to exclude outlier glucose readings, we only evaluate treatment effects between the 2.5 and 97.5 percentiles.

**ADD MALTS vs Difference of Means** Because the data we analyze is from a randomized experiment, the difference in mean quantile functions will be an unbiased estimate of average treatment effects. We compare ADD MALTS’ estimated ATE to the difference-in-means-estimated ATE (DIME ATE). As seen in Figure 6, the two estimates are not significantly different from one another (the ADD MALTS ATE is within the 95% confidence interval for the DIME ATE). Also, there is a very marginal difference between the point-estimates: the DIME ATE also ranges between -2 and 2, and the difference between the DIME and ADD MALTS’ ATEs range between -1 and 1. The fact that ADD MALTS’ estimated ATE is so close to the DIME ATE validates ADD MALTS’ ability to estimate average treatment effects in real-world settings.

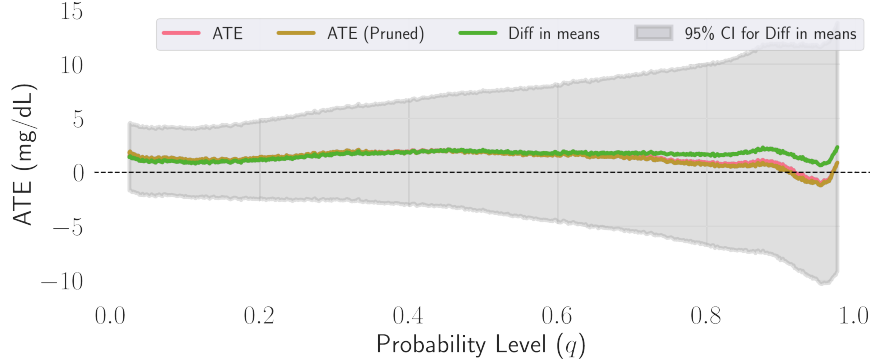


Figure 6: We compare the difference-in-means-estimated ATE (green) to the ADD MALTS’ estimated ATE before and after pruning bad matched groups (pink, gold respectively). The grey interval represents 95% confidence interval for the difference-in-means-estimated ATE.

## Appendix E Additional Synthetic Experiments

In this section, we provide more details on the CATE estimation experiments and the positivity violation experiment. We also demonstrate how ADD MALTS’ CATE estimation performance is insensitive to the number of nearest neighbors.

### E.1 CATE Estimation Experimental Details

In this section, we provide more details on CATE estimation experiments in the main paper.

Name	$\mu$	$\sigma$
Variance	$10 + x_{i,0} + 2x_{i,1} + \varepsilon_i$	$ 10 + x_{i,0} + 2x_{i,1} + 10T_i + \varepsilon_i $
Linear	$10 + x_{i,0} + 2x_{i,1} + 10T_i + \varepsilon_i$	1
Complex	$10 \sin(\pi x_{i,0} x_{i,1}) + 20(x_{i,2} - 0.5)^2 + 10x_{i,3} + 5x_{i,4} + T_i \{7 + x_{i,2} \cos(\pi x_{i,0} x_{i,1})\} + \varepsilon_i$	$ 10 + x_{i,0} + 2x_{i,1} + \varepsilon_i $
Dist Cov	$10 \sin\left(\pi \left[\int_0^1 F_{x_{i,0}}^{-1}(q) dq\right] x_{i,1}\right) + 20(x_{i,2} - 0.5)^2 + 10x_{i,3} + 5x_{i,4} + T_i \left\{7 + x_{i,2} \cos\left(\pi \left[\int_0^1 F_{x_{i,0}}^{-1}(q) dq\right] x_{i,1}\right)\right\} + \varepsilon_i$	1

Table 3: In our simulations, the distributional outcomes are truncated normal distributions (truncated at  $\pm 3$  standard deviations from the mean). The table describes the functions used to generate the mean and variance of each outcome. In the “Linear,” “Variance,” and “Complex” DGPs, we respectively sample 2, 2, and 6 *scalar* covariates independently and identically from  $Uniform[-1, 1]$ . The last simulation (Dist Cov) has a distributional covariate  $F_{x_{i,0}}$ , which we use by taking the integral of its cumulative distribution function as a covariate, and 9 scalar covariates.

We consider four data generative processes (DGPs). In each DGP, we generate our distributional outcomes as truncated normal distributions (truncated at  $\pm 3$  standard deviations from the mean). Table 3 describes the functions used to generate the means and variances of the truncated normal

outcomes for each DGP. In the “Linear,” “Variance,” and “Complex” DGPs, we respectively sample 2, 2, and 6 *scalar* covariates independently and identically from  $Uniform[-1, 1]$ . The propensity score models are linear:  $\mathbb{P}(T_i = 1|F_{\mathbf{x}_i}) = \text{expit}(x_{i,0} + x_{i,1})$ .

The “Dist Cov” simulation instead has nine scalar covariates and a distributional covariate  $F_{x_{i,0}}$ .  $F_{x_{i,0}}$  is uniformly sampled to be any uniform distribution between  $[-1, 0]$  and  $[0, 1]$ . We use the integral of the quantile function as a feature to generate outcomes. Because baseline methods can only handle scalar covariates, we instead provide them the preprocessed covariate, the area under the quantile function. Here, the propensity score model is also linear but depends on this processed distribution:  $\mathbb{P}(T_i = 1|F_{\mathbf{x}_i}) = \text{expit}\left(\left[\int_0^1 F_{x_{i,0}}^{-1}(q) dq\right] + x_{i,2}\right)$ .

For each DGP, we consider 1500 units, using 67% in the training set and 33% to estimate CATEs. Also, each DGP has at least five irrelevant covariates. We repeat each experiment 100 times and evaluate CATE estimation performance using the Integrated Relative Error (%) =  $100 \cdot \int_0^1 \left| \frac{\hat{\tau}(q) - \tau(q)}{\tau(q)} \right| dq$ .

We compare ADD MALTS to the following baselines: **Lin PSM** and **RF PSM** represent propensity score matching fit with linear and random forest models, respectively. We use a cross-validated  $\ell_1$ -regularized logistic regression to train the linear propensity score model. We use the default settings in `sklearn`’s random forest implementation to train the random forest propensity score model. **FT** and **FRF** represent decision tree and random forest methods for functional outcomes (Qiu et al., 2022); we use a depth bound of 5 to train the decision tree and 100 trees of depth bound 20 to train our forest<sup>2</sup>. **LR** represents outcome regression fit at each quantile with a linear regression (Lin et al., 2023); **LR + Lin PS** and **LR + RF PS** represent augmented inverse propensity weighting methods combining the linear outcome regression with linear and random forest propensity score models, respectively. Here, we use an unregularized linear regression for the outcome model, as in Lin et al. (2023), and use the same propensity score model configurations as in the propensity score matching methods. Table 4 details the package dependencies of each method that we used for implementation.

### E.1.1 Timing Results

In this section, we evaluate ADD MALTS’ running time. ADD MALTS had median run-time across all Monte Carlo iterations less than 10 minutes, from fitting the distance metric to estimating CATEs (see Figure 7). The Linear and Variance DGPs had the lowest median runtime, less than three minutes per iteration.

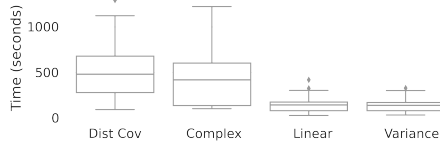


Figure 7: Each boxplot displays the time in seconds (y-axis) it took for ADD MALTS to run, from fitting the distance metric to estimating CATEs, for each DGP in Table 3 across 100 Monte Carlo iterations.

---

**Algorithm 1** Mixture-beta outcomes DGP

---

```

for  $i = 1, \dots, 1000$  do
   $X_{i,1}, \dots, X_{i,10} \sim \text{Unif}[-1, 1]$  ▷ Generate scalar covariates
   $T_i \sim \text{Bern}(\text{expit}(X_{i,1} + X_{i,2}))$  ▷ Assign treatment
   $\alpha(t_i) = 5 + 10 \sin(\pi X_{i,1} X_{i,2})^2 + 20(X_{i,3} - 0.5)^2 + 10X_{i,4} + 5X_{i,5} + t_i 10X_{i,3} \cos(\pi X_{i,1} X_{i,2})^2 +$ 
   $\varepsilon_{Y_i}, \varepsilon_{Y_i} \sim \mathcal{N}(0, 1)$  ▷ Generate parameter that will control Beta distribution's shape
  for  $j = 1, \dots, 1001$  do
     $Z_{i,j} \sim \text{Bern}(1/4)$  ▷ Sample which mixture of the outcome this observation will come from
     $Y_{i,j}(0) \sim \text{Beta}(2\alpha(0), 8\alpha(0))^{Z_{i,j}} \text{Beta}(8\alpha(0), 2\alpha(0))^{1-Z_{i,j}}$  ▷ Generate sample from the
    control potential outcome
     $Y_{i,j}(1) \sim \text{Beta}(2\alpha(1), 8\alpha(1))^{1-Z_{i,j}} \text{Beta}(8\alpha(1), 2\alpha(1))^{Z_{i,j}}$  ▷ Generate sample from the
    treated potential outcome
  end for
   $F_{Y_i(0)}^{-1}(q) = \min\{y : \frac{1}{1001} \sum_{j=1}^{1001} \mathbb{1}[Y_{i,j}(0) \leq y] \geq q\}$  ▷ Generate the control potential
  outcome's quantile function
   $F_{Y_i(1)}^{-1}(q) = \min\{y : \frac{1}{1001} \sum_{j=1}^{1001} \mathbb{1}[Y_{i,j}(1) \leq y] \geq q\}$  ▷ Generate the treated potential
  outcome's quantile function
   $\tau(q|F_{\mathbf{x}_i}) = F_{Y_i(1)}^{-1}(q) - F_{Y_i(0)}^{-1}(q)$  ▷ Calculate the true CATE
   $F_{Y_i}^{-1}(q) = T_i F_{Y_i(1)}^{-1}(q) + (1 - T_i) F_{Y_i(0)}^{-1}(q)$  ▷ Calculate the observed outcome
end for return  $\mathcal{S}_n = \left\{ (X_{i,1}, \dots, X_{i,10}), T_i, F_{Y_i}^{-1} \right\}_{i=1}^n$  ▷ Return observations

```

---

## E.2 ADD MALTS' CATE Estimation Performance on Another DGP

In this section, we also evaluate ADD MALTS' CATE estimation performance when the outcome distribution is not a truncated normal distribution. In this experiment, we generate our outcomes as a mixture of two-beta distributions (as seen in Algorithm 1).

The DGP has 10 scalar covariates. We then assign treatment using a simple, linear propensity score model using two of the covariates. Our outcome is a mixture of two Beta distributions whose parameters are controlled by the term  $\alpha(t_i)$ ;  $\alpha(t_i)$  is a combination of complex quadratic and trigonometric terms. When  $t_i = 1$ , the mixture proportions flip so that more mass is concentrated

---

<sup>2</sup>The choice of 100 trees comes from the default setting in `sklearn`'s implementation of random forests. While `sklearn` has no depth bound, the functional outcome tree code we wrote ran into memory issues when the depth bound was greater than 20

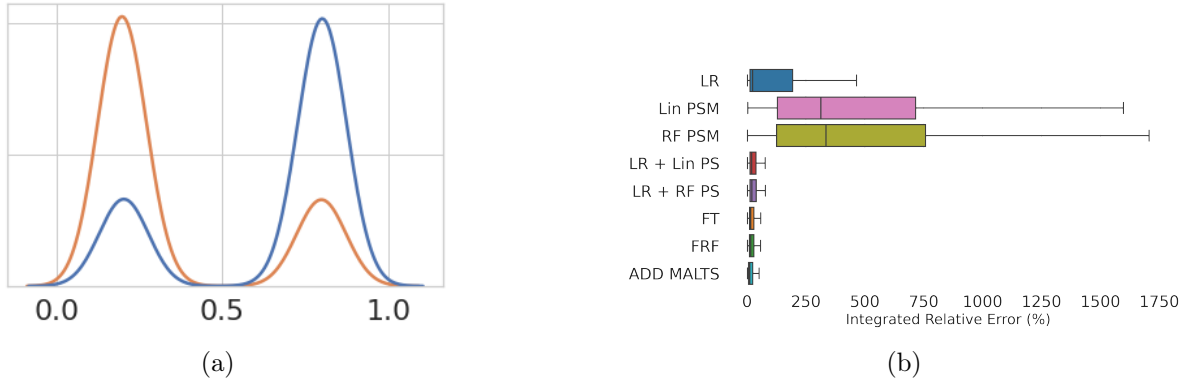


Figure 8: Subplot (a) displays the densities of the treated (blue) and control (orange) potential outcomes generated from this data generating process. Subplot (b) displays the integrated relative error (y-axis) for the various baselines and ADD MALTS for the DGP described in Algorithm 1.

in the upper end of the distribution’s support; additionally, when  $t_i = 1$ , the Beta distribution’s parameter increases by another complex interaction of trigonometric terms, causing the variance of each mixture component to shrink. We construct each unit’s observed quantile function using the outcome observations associated with unit  $i$ ’s treatment status:  $y_{i,1}(t_i), \dots, y_{i,1001}(t_i)$ . In our DGP, we have 5 relevant covariates and 5 irrelevant covariates. We also use a 60/40 train/estimation split to estimate conditional average treatment effects. As seen in Figure 8, ADD MALTS does at least as well as the other methods in this complex DGP.

### E.3 Positivity Violation Experimental Details

In this section, we provide more details on the Positivity Violations experiment in Section 4.2. While we describe the DGP we considered in the main paper, we offer more details on the implementation of the baselines and ADD MALTS in this section. We have two baseline methods: a linear propensity score model fit using cross-validation (using the default LASSO logistic regression cross-validation parameters in `sklearn`) and a random forest propensity score model fit using cross-validation (cross-validating over the number of trees: 20, 50, 100, 200). We use `sklearn`’s cross-validation implementations of both these techniques to find the best parameters.

**Propensity score flagging methods and their relationship to the propensity score thresholds** We evaluate how sensitive RF PS and Lin PS are to changes in the propensity score thresholds when flagging units as being in positivity violation regions. As evidenced in Figure 9, ADD MALTS outperforms the other methods with various propensity score thresholds. Furthermore, the qualitative inspection of nearest neighbor sets using ADD MALTS offers a layer of fidelity unachievable by the propensity score methods that produce uninterpretable matched groups (see Parikh et al.

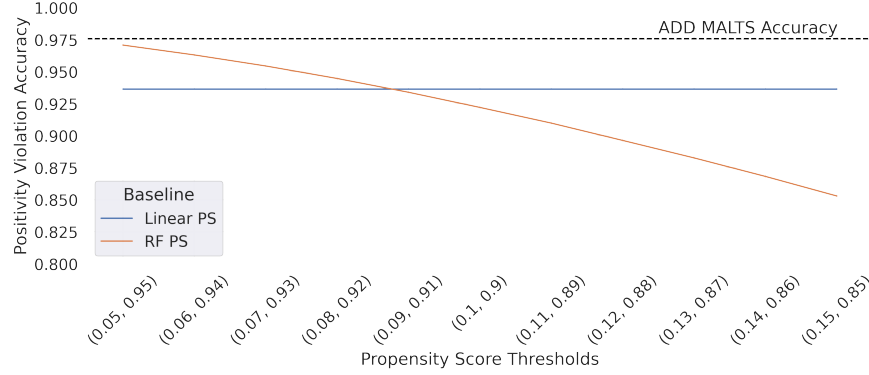


Figure 9: The x-axis represents the thresholds used for flagging regions of the covariate space that may suffer from a positivity violation using the propensity score methods. The y-axis represents each method’s accuracy in flagging units that are in positivity violation regions. The orange line represents the accuracy for the random forest estimated propensity score while the blue line represents the accuracy for the propensity score estimated with a linear model. The black, dotted line represents the accuracy when using ADD MALTS’ diameter to flag units.

(2022) for more details on this comparison).

#### E.4 Insensitivity to the Number of Nearest Neighbors

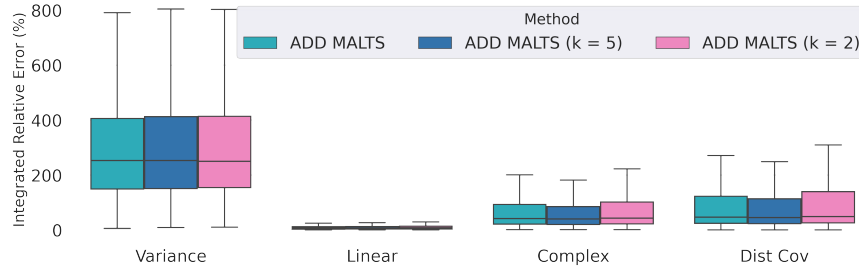


Figure 10: The figure displays the integrated relative error (%) on the y-axis of each DGP (x-axis) for ADD MALTS estimators with different numbers of nearest neighbors (10 in turquoise, 5 in navy blue, 2 in pink).

In this section, we demonstrate that ADD MALTS’ CATE estimation performance is not affected by the choice of the number of nearest neighbors. Figure 10 displays the integrated relative error (%) on the y-axis of each DGP (x-axis) for ADD MALTS estimators with different numbers of nearest neighbors (10, 5, 2). All box-plots overlap, with the largest difference being a deviation of about 50% IRE between  $k = 5$  and  $k = 2$  in the “Dist Cov” simulation. A 50% difference in IRE is marginal compared to the range of values seen in Figure 2 of 0-1500%.

## E.5 Computational Resources

All experiments for this work were performed on an academic institution’s cluster computer. We used up to 40 machines in parallel, selected from the specifications below:

- 2 Dell R610’s with 2 E5540 XeonProcessors (16cores)
- 10 Dell R730’s with 2 Intel Xeon E5-2640 Processors (40 cores)
- 10 Dell R610’s with 2 E5640 Xeon Processors (16 cores)
- 10 Dell R620’s with 2 Xeon(R) CPU E5-2695 v2’s (48 cores)
- 8 Dell R610’s with 2 E5540 Xeon Processors (16cores)

We did not use GPU acceleration for this work.

Method	Code Dependencies
Lin PSM	Pedregosa et al. (2011); Harris et al. (2020)
RF PSM	Pedregosa et al. (2011); Harris et al. (2020)
FT	Pedregosa et al. (2011); Harris et al. (2020); pandas development team (2020)
FRF	Pedregosa et al. (2011); Harris et al. (2020); pandas development team (2020)
LR	Pedregosa et al. (2011); Harris et al. (2020)
LR + Lin PS	Pedregosa et al. (2011); Harris et al. (2020)
RF + Lin PS	Pedregosa et al. (2011); Harris et al. (2020)
ADD MALTS	Harris et al. (2020); Virtanen et al. (2020); pandas development team (2020)

Table 4: A table describing which libraries each baseline in the CATE estimation experiment depended on.