COUNTERFACTUAL DELAYED FEEDBACK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

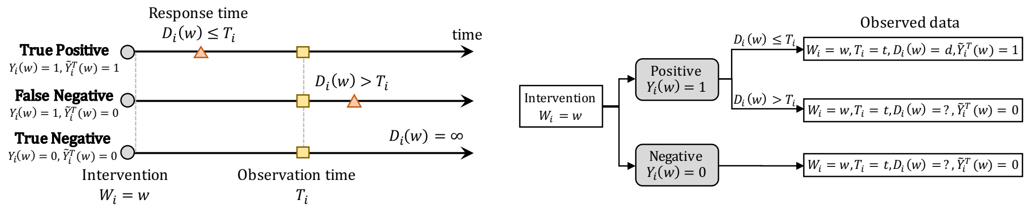
Estimation of heterogeneous treatment effects has gathered much attention in recent years and has been widely adopted in medicine, economics, and marketing. Previous studies assumed that one of the potential outcomes of interest could be observed timely and accurately. However, a more practical scenario is that treatment takes time to produce causal effects on the outcomes. For example, drugs take time to produce medical utility for patients and users take time to purchase items after being recommended, and ignoring such delays in feedback can lead to biased estimates of heterogeneous treatment effects. To address the above problem, we study the impact of observation time on estimating heterogeneous treatment effects by further considering the potential response time that potential outcomes have. We theoretically prove the identifiability results and further propose a principled learning approach, known as CFR-DF (Counterfactual Regression with Delayed Feedback), to simultaneously learn potential response times and potential outcomes of interest. Results on both simulated and real-world datasets demonstrate the effectiveness of our method.

1 INTRODUCTION

Heterogeneous treatment effects (HTE) estimation using observational data is a fundamental problem that applies to a wide variety of areas (Alaa & Van Der Schaar, 2017; Alaa et al., 2017; Hannart et al., 2016; LaLonde, 1986; Shalit et al., 2017). For example, in precision medicine, physicians decide drug allocation by the treatment effect of the patient on the drug (Jaskowski & Jaroszewicz, 2012). In online markets, the causal effect of recommending an item on a user's purchase behavior is used for personalized recommendations (Schnabel et al., 2016). Unlike using observed outcomes to make decisions, HTE accounts for variations in both factual outcomes and counterfactual outcomes among individuals or subgroups. The challenge lies in accurately estimating HTE due to the unobserved counterfactual outcomes with alternative treatment (Holland, 1986).

Many methods have been proposed to estimate HTE from observational data. For instance, representation learning-based approaches learn a covariate representation that is independent of the treatment to overcome the covariate shift between the treatment and control groups (Johansson et al., 2016; Shalit et al., 2017; Shi et al., 2019; Yao et al., 2018). The tree-based approach generalizes Bayesian inference and random forest methods for nonparametric estimation (Chipman et al., 2010; Wager & Athey, 2018). The generative model-based approaches use the widely adopted variational autoencoder and generative adversarial network to generate individual counterfactual outcomes (Louizos et al., 2017; Yoon et al., 2018). These studies have also been extended to continuous treatment scenarios (Bica et al., 2020; Nie et al., 2021; Schwab et al., 2018; 2020).

Existing methods require that one of the potential outcomes of interest be observed timely and accurate. However, interventions on individuals usually do not affect outcomes of interest immediately, and treatment takes time to produce causal effects on the outcomes. For example, drugs take time to produce medical utility for patients, with the long-term prognosis as the outcome of interest, which benefits the treatment decision from the physicians. In online markets, a recommendation algorithm focuses on whether or not the user will eventually purchase, but users take time to purchase items after being recommended (Chapelle, 2014), which poses a critical challenge in practice: as in Figure 1(a), if the observation window is too short, some samples will be incorrectly marked as negative whose conversion will occur in the future; but if it is too long, the recommendation algorithm will not be able to guarantee its timely availability (Yoshikawa & Imai, 2018). In summary, ignoring such delays in outcome response can lead to biased estimates of HTE.



(a) Three types of delayed response scenarios. (b) Observed data with various potential outcomes.

Figure 1: Illustrations for false negative (left) and observed data format (right) under delayed response.

In this paper, we first formalize the HTE estimation problem in the presence of delayed response. In contrast to previous studies that only considered the effect of treatment on outcome, we also consider potential response times with different treatments, since treatment may affect response time, e.g., users who receive item recommendations purchase more quickly. Therefore, as in Figure 1(a), given the treatment w for an individual, even if the eventual outcome of interest $Y(w)$ is positive, e.g., the user will eventually purchase the item, we can only observe the true positive conversion ($Y(w) = 1, \tilde{Y}(w) = 1$) when the potential response time is less than the observation time ($D(w) \leq T$), while observing the false negative outcome ($Y(w) = 1, \tilde{Y}(w) = 0$) vise versa. Instead, when the eventual outcome $Y(w)$ is negative, e.g., the user never purchases the item, then we observe the negative outcome ($\tilde{Y}(w) = 0$) regardless of the observation time. Figure 1(b) illustrates the format of the observed data, which comes with an additional challenge, that is, we could not obtain the exact value of the response time if the positive feedback did not occur before the observation time.

To address the above issues, we study the impact of observation time on estimating heterogeneous treatment effects by further considering the potential response time that potential outcomes have. Theoretically, we prove the eventual potential outcomes are identifiable in the whole population, which is essential for treatment allocation. For subgroups in which individuals always have positive eventual outcomes regardless of treatment, we also show the identifiability of potential response times, which quantifies the causal effect of treatment on response times. Using the eventual outcomes as hidden variables, we reconstruct the posterior distribution of a delayed response and provide explicit solutions to estimate the parameters of interest within a modified EM algorithm. Furthermore, we propose a principled learning approach that extends counterfactual regression (CFR) to delayed feedback outcomes, named CFR-DF, to simultaneously predict potential outcomes and potential response times. Finally, we discuss the importance of this work for policy learning and validate the effectiveness of the proposed method on both synthetic and real-world datasets.

The main contributions of this paper are summarized as follows:

- We formalize the HTE estimation problem with delayed response, in which treatment takes time to produce a causal effect on the outcome.
- We theoretically prove the eventual potential outcome is identifiable, and also show the identifiability of potential response times on the always-positive stratum.
- We propose a principled learning algorithm, called CFR-DF, that utilizes the EM algorithm to estimate both eventual potential outcomes and potential response times.
- We perform extensive experiments on both synthetic and real-world datasets to show the effectiveness of the proposed approach in estimating HTE with delayed responses.

2 Heterogeneous Treatment Effect with Delayed Response

2.1 Notation and Setup

In this paper, we consider the case of binary treatment. Suppose a simple random sample of n units from a super population $\mathbb{P}$, for each unit i, the covariate and the assigned treatment are denoted as $X_i \in \mathcal{X} \subset \mathbb{R}^m$ and $W_i \in \mathcal{W} = \{0, 1\}$, where $W_i = 1$ means receiving the treatment and $W_i = 0$ means not receiving the treatment, respectively. Different from the previous problem setup in both standard HTE estimation (Johansson et al., 2016; Shalit et al., 2017; Shi et al., 2019; Yao et al., 2018)

Table 1: The units are divided into four strata based on the joint potential outcomes $(Y(0), Y(1))$.

Group	$Y(0)$	$Y(1)$	$D(0)$	$D(1)$	Preferred treatment
PP	1	1	✓	✓	Depends on $\tau_D(x)$
NP	0	1	∞	✓	Treatment ($W = 1$)
PN	1	0	✓	∞	Control ($W = 0$)
NN	0	0	∞	∞	Either ($W = 0$ or 1)

and recent time-to-event studies related to survival analysis (Gupta et al., 2023; Chapfuwa et al., 2021; Curth et al., 2021), we consider the response time from the imposing treatment to producing influence on the outcome. Specifically, let $Y_i \in \mathcal{Y} = \{0, 1\}$ be the binary outcome at the eventual time, e.g., whether a user will eventually purchase, as the primary outcome of interest, and we call unit with $Y_i = 1$ as a positive sample. Without loss of generality, the time at which the treatment W_i is imposed on unit i is taken as the start time, let D_i be the response time for individuals with $Y_i = 1$ to produce positive feedback, and we set $D_i = \infty$ for individuals with $Y_i = 0$. As shown in Figure 1(a), given an observation time T_i, we see a positive feedback at T_i, denoted as $\tilde{Y}_i^T = 1$, if and only if individual i is a positive sample $Y_i = 1$ with the response time $D_i \leq T_i$, and marked as *true positive*. However, for some other positive samples with $Y_i = 1$, we would see false negative feedback $\tilde{Y}_i^T = 0$ at the observation time T_i, when the response time is greater than the observation time, i.e., $D_i > T_i$, and marked as *false negative*. For samples that never yield positive outcomes, we observe negative feedback $\tilde{Y}_i^T = 0$ for all observation times T_i, and marked as *true negative*.

To study the effect of treatment on the eventual outcome and the response time, we adopt the potential outcome framework (Rubin, 1974; Neyman, 1990) in causal inference. Specifically, let $Y_i(0)$ and $Y_i(1)$ be the eventual outcome of unit i had this unit receive treatment $W_i = 0$ and $W_i = 1$, respectively. In addition, since treatment may have an effect on the response time, e.g., users purchase more quickly when receiving ads about an item, we denote $D_i(0)$ and $D_i(1)$ be the potential response time had unit i receive treatment $W_i = 0$ and $W_i = 1$, respectively. Therefore, given an observation time T_i, the corresponding potential outcomes $\tilde{Y}_i^T(0)$ and $\tilde{Y}_i^T(1)$ can be analogously defined. Since each unit can be only assigned with one treatment, we always observe the corresponding outcome to be either $\tilde{Y}_i^T(0)$ or $\tilde{Y}_i^T(1)$, but not both, which is also known as the fundamental problem of causal inference (Holland, 1986; Morgan & Winship, 2015). However, one should note that similar conclusions no longer hold for the eventual potential outcomes $(Y_i(0), Y_i(1))$ and the potential response times $(D_i(0), D_i(1))$, as we cannot observe the exact eventual outcome as well as the response time due to the limited observation time.

We assume that the observation for unit i is $\tilde{Y}_i^T = (1 - W_i)\tilde{Y}_i^T(0) + W_i\tilde{Y}_i^T(1)$. In other words, the observed outcome at time T_i is the potential outcome corresponding to the assigned treatment, which is also known as the consistency assumption in the causal literature. We assume that the stable unit treatment value assumption (STUVA) assumption holds, i.e., there should not be alternative forms of treatment and interference between units. Furthermore, we assume the positivity of treatment assignment, i.e., $\eta < \mathbb{P}(W_i = 1|X_i = x) < 1 - \eta$, where η is a constant between 0 and $1/2$.

We summarize the observed data formats in Figure 1(b), with the following three cases.

- True positive ($Y_i(w) = 1$, $\tilde{Y}_i^T(w) = 1$) with observed ($W_i = w, D_i(w) = d \leq T_i, \tilde{Y}_i^T(w) = 1$);
- False negative ($Y_i(w) = 1$, $\tilde{Y}_i^T(w) = 0$) with observed ($W_i = w, T_i = t, \tilde{Y}_i^T(w) = 0$);
- True negative ($Y_i(w) = 0$, $\tilde{Y}_i^T(w) = 0$) with observed ($W_i = w, T_i = t, \tilde{Y}_i^T(w) = 0$),

which leads to an additional challenge due to one cannot distinguish between *false negative* and *true negative* directly from the observed data ($W_i = w, T_i = t, \tilde{Y}_i^T(w) = 0$).

2.2 Parameters of Interest

We consider two meaningful parameters of interest in the following. For simplification, we drop the subscript i for a generic unit hereafter. First, unlike previous studies that focused on the HTE of treatment on current observed outcomes, i.e., $\tau^T(x) = \mathbb{E}[\tilde{Y}^T(1) - \tilde{Y}^T(0) \mid X = x]$, we focused on the HTE of treatment on the eventual outcomes, i.e., $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$. Notably, the latter poses two challenges: first, the confounding bias introduced by covariates, which is similar to previous studies; second, how to recover the eventual outcome Y of interest from the observed outcome $\tilde{Y}^T$ at time T. When the observation time T is sufficiently long to exceed the response time D for all individuals, the proposed causal estimand $\tau(x)$ degenerates to $\tau^T(x)$.

Next, we show that individuals can be divided into four strata by considering the joint potential outcomes $(Y(0), Y(1))$, as shown in Table 1, and named as the *always-positive* stratum, *useful treatment* stratum, *harmful treatment* stratum, and *always-negative* stratum accordingly. From a policy learning perspective, it is clear that treatment should be given and not given to individuals in *useful treatment* stratum and *harmful treatment* stratum, respectively. For individuals in the *always-negative* stratum, for example, users who will never purchase or patients who will always be cured regardless of treatment, either of the treatments is reasonable and results in no difference. When considering individuals in the *always-positive* stratum, despite having both $Y(0) = 1$ and $Y(1) = 1$ for the eventual outcomes, it is meaningful to study the HTE of the treatment on the response times. Formally, the causal estimand of interest is $\mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$. For the other three strata, since there exists a treatment w such that $Y(w) = 0$, the corresponding response time can be regarded as $D(w) = \infty$, resulting in HTE of treatment on response time being ill-defined.

We summarize the causal estimand of interest as follows.

- HTE on the eventual outcome: $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$;
- HTE on the response time: $\tau_D(x) = \mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$.

2.3 Identifiability Results

We then discuss the identifiability of the causal parameters of interest in Section 2.2. We adopt and refer to the following assumptions.

Assumption 1 (Unconfoundedness). $W \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1)) \mid X$ *for all* $t > 0$.

Assumption 2 (Time Independence). $T \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1), W) \mid X$ *for all* $t > 0$.

Assumption 3 (Time Sufficiency). $\inf\{d : F_D^{(w)}(d \mid Y(w) = 1, X) = 1\} < \inf\{t : F_T(t) = 1\}$ *for* $w = 0, 1$*, where* $F(\cdot)$ *is the cumulative distribution function (cdf).*

Assumption 4 (Monotonicity). $Y(0) \leq Y(1)$.

Assumption 5 (Principal Ignorability). $(W, Y(w)) \perp\!\!\!\perp D(1 - w) \mid Y(1 - w), X$ *for* $w = 0, 1$.

Among them, unconfoundedness is also known as no unmeasured confounders assumption as it holds if all variables that affect both treatment and potential outcomes are included in X. Time independence holds since the observation occurs after the treatment, and the observation does not affect the potential response times $D(w)$ and the potential outcomes $\tilde{Y}^t(w)$ at a given time $t > 0$ for $w = 0, 1$. Time Sufficiency means that we need a subset of individuals (not all) with observed outcomes $\tilde{Y} = 1$ to identify eventual potential outcomes, which is a necessary condition for studying survival analysis. Monotonicity assumption is plausible in many applications when the effect of the decision on the outcome is non-negative for all individuals, e.g., the drug is not harmful to the patient or recommendations do not have a negative effect on user purchases. Principal Ignorability requires that the expectations of the potential outcomes do not vary across principal strata conditional on the covariates. It is widely used in applied statistics (Imai & Jiang, 2020; Ben-Michael et al., 2022).

We next provide the identifiability results of three causal parameters (see Appendix A.2.1 for proofs).

Theorem 1. *Under Assumptions 1-3, the HTE on the eventual outcome* $\tau(x)$ *is identifiable.*

To identify the HTE of treatment on potential response times in the *always-positive* stratum, we introduce the monotonicity assumption to identify the probability of belonging to this stratum.

Lemma 1. *Under Assumptions 1-4,* $\mathbb{P}(Y(0) = 1, Y(1) = 1 \mid X = x)$ *is identifiable.*

Following the previous studies (Imai & Jiang, 2020; Ben-Michael et al., 2022; Jiang et al., 2022), we assume principal ignorability holds to identify the HTE of treatment on potential response times in the *always-positive* stratum. Under all of the above assumptions, $\tau_D(x)$ is also identifiable.

Theorem 2. *Under Assumptions 1-5, the HTE on the response time in the always-positive stratum* $\tau_D(x) = \mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$ *is identifiable.*

3 CFR-DF: COUNTERFACTUAL REGRESSION WITH DELAYED FEEDBACK

In this section, we propose a principled learning approach to perform **C**ounter**F**actual **R**egression with **D**elayed **F**eedback on outcomes, named CFR-DF. Specifically, CFR-DF consists of two sets of models to predict the eventual potential outcomes, i.e., $\mathbb{P}(Y(0) = 1 \mid X = x)$ and $\mathbb{P}(Y(1) = 1 \mid X = x)$ and the potential response times, i.e., $\mathbb{P}(D(0) = d \mid X = x, Y(0) = 1)$ and $\mathbb{P}(D(1) = d \mid X = x, Y(1) = 1)$, respectively, the former of which can be flexibly exploited from previous HTE estimation methods in the following framework, and we take the widely used counterfactual regression (CFR) (Shalit et al., 2017) for illustration purpose.

Recall that in Figure 1(b), we show two possible observed data formats. On the one hand, the probability of observing positive feedback $\tilde{Y}^T = 1$ with response time $D = d$ at time $T = t > d$:

$$
\begin{aligned}
&p(\tilde{Y}^T = 1, D = d \mid X = x, W = w, T = t) = p(Y = 1, D = d \mid X = x, W = w) \\
&= \mathbb{P}(Y(w) = 1 \mid X = x, W = w) p(D(w) = d \mid X = x, W = w, Y(w) = 1) \\
&= \mathbb{P}(Y(w) = 1 \mid X = x) p(D(w) = d \mid X = x, Y(w) = 1),
\end{aligned}
$$

where the first equality follows from time independence, the second equality follows from the consistency assumption, and the last equality follows from the unconfoundedness assumption.

On the other hand, by the law of total probabilities, and again using the conditional independence of observation time, the probability of not having observed positive feedback at time $T = t > d$ is:

$$
\begin{aligned}
&\mathbb{P}(\tilde{Y}^T = 0 \mid X = x, W = w, T = t) \\
&= \mathbb{P}(Y = 0 \mid X = x, W = w) \mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 0) \\
&+ \mathbb{P}(Y = 1 \mid X = x, W = w) \mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 1),
\end{aligned}
$$

where $\mathbb{P}(Y = 0 \mid X = x, W = w)$ is equivalent to $\mathbb{P}(Y(w) = 0 \mid X = x)$ by unconfoundedness assumption, with similar result holds for $\mathbb{P}(Y = 1 \mid X = x, W = w)$. In addition, we have $\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 0) = 1$, due to eventual outcome $Y = 0$ implies $\tilde{Y}^t = 0$ for all $t > 0$. Next we focus on the last item $\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 1)$.

By noting the equivalence between $(\tilde{Y}^t(w) = 0, Y(w) = 1)$ and $(D(w) > t, Y(w) = 1)$, we have:

$$
\begin{aligned}
&\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 1) = \mathbb{P}(D(w) > t \mid X = x, Y(w) = 1) \\
&= \int_t^\infty p(D(w) = u \mid X = x, Y(w) = 1) du.
\end{aligned}
$$

With the above results, we have the probability of $\tilde{Y}^T = 0$ at time $T = t$ is:

$$
\begin{aligned}
&\mathbb{P}(\tilde{Y}^T = 0 \mid X = x, W = w, T = t) = \mathbb{P}(Y(w) = 0 \mid X = x) \\
&+ \mathbb{P}(Y(w) = 1 \mid X = x) \int_t^\infty p(D(w) = u \mid X = x, Y(w) = 1) du,
\end{aligned}
$$

which can be represented by two sets of models in CFR-DF.

Different from CFR, an essential challenge is that we cannot observe the eventual outcomes Y, which results in the unavailability to directly fit the potential outcomes of interest $\mathbb{P}(Y(w) = 0 \mid X = x)$ and $\mathbb{P}(Y(w) = 1 \mid X = x)$ from the observed data. To address this problem, we treat the eventual potential outcomes as latent variables, and estimate the parameters of interest using a modified EM algorithm as below, which addresses both the confounding bias and the missing eventual outcomes.

Expectation Step. For a given data point (x_i, w_i, t_i, y_i^t), we need to compute the posterior probability of the hidden variable $p_i := \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i, W = w_i, T = t_i, \tilde{Y}^T = y_i^t)$. If positive feedback $y_i^t = 1$ is observed at time $T = t$, then it is obvious that $p_i = 1$ for unit i. Alternatively, if $y_i^t = 0$ is observed at time t for individual i, then the posterior probability p_i can be expressed as:

$$
\begin{aligned}
p_i &= \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i, W = w_i, T = t_i, \tilde{Y}_i^T = 0) \\
&= \frac{\mathbb{P}(\tilde{Y}_i^T(w_i) = 0 \mid X = x_i, Y_i(w_i) = 1, T = t_i) \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i)}{\mathbb{P}(\tilde{Y}_i^T = 0 \mid X = x_i, W = w_i, T = t_i)},
\end{aligned}
$$

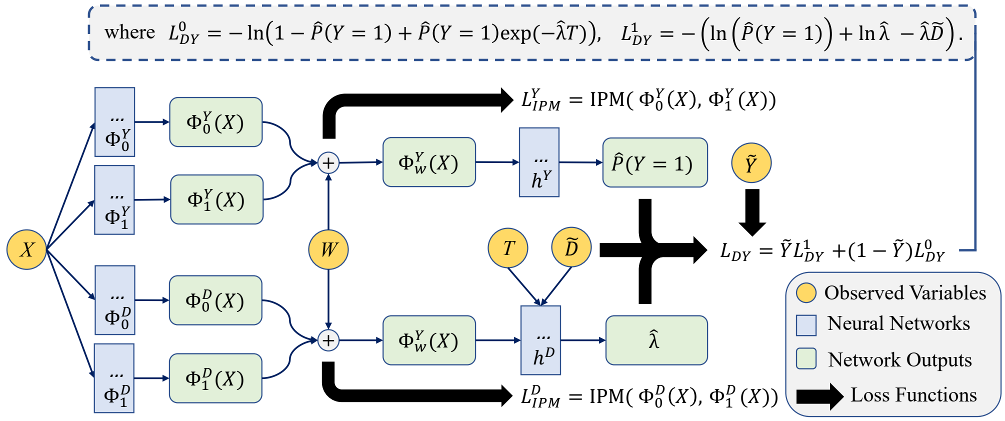


Figure 2: Overview of CFR-DF Architecture. For the representation block, we use multi-layer neural networks Φ with ELU activation function to learn representation and each network has two/three layers with m_X units, respectively. Then, we use a single-layer network h^Y with Sigmoid activation to achieve $\hat{P}(Y=1)$ and a single-layer network h^D with SoftPlus sigmoid activation to achieve $\hat{\lambda}$.

which can be calculated from the maximization step of the models in CFR-DR in the following.

Maximization Step. Given the hidden variable values p_i computed from the E step, let $S = s_i$ denote $(X = x_i, W = w_i, T = t_i)$, we maximize the expected log-likelihood during the M step:

$$\sum_i p_i \log \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i) + \sum_i (1-p_i)\log(1 - \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i))$$
$$+ \sum_{i:\tilde{y}_i^t = 1} \log p(D_i(w_i) = d_i \mid X = x_i, Y_i(w_i) = 1)$$
$$+ \sum_{i:\tilde{y}_i^t = 0} p_i \log \int_{t_i}^{\infty} p(D(w_i) = u \mid X = x_i, Y_i(w_i) = 1) du,$$

where *the eventual potential outcome model* $\mathbb{P}(Y(w) = 1 \mid X = x)$ *and the potential response time model* $p(D(w) = d \mid X = x, Y(w) = 1)$ *can be optimized independently.* Due to space limitations, the computation details of parametric and non-parametric EM models are deferred to Appendix A.2.2.

Let $h^Y(\Phi^Y(x), w)$ be the prediction model for the eventual potential outcomes $\mathbb{P}(Y(w) = 1 \mid X = x)$, and $h^D(\Phi^D(x), w, d)$ be the prediction model for the potential response times $p(D(w) = d \mid X = x, Y(w) = 1)$, where $\Phi^Y : \mathcal{X} \to \mathcal{R}^Y$ and $\Phi^D : \mathcal{X} \to \mathcal{R}^D$ are the covariate representations, $\mathcal{R}^Y$ and $\mathcal{R}^D$ are the representation spaces, and $h^Y : \mathcal{R}^Y \times \{0,1\} \to \mathcal{Y}$ and $h^D : \mathcal{R}^D \times \{0,1\} \times \mathbb{R}^+ \to \mathbb{R}^+$ are the prediction heads, respectively. Inspired by CFR (Shalit et al., 2017), we take the Integral Probability Metric (IPM) distance induced by the representations as a penalty term, to control the generalization error caused by covariate shift between the treatment and control group.

Given the posterior probabilities p_i computed from the E step, we train the eventual potential outcome model by minimizing the derived negative log-likelihood in the M step with the IPM distance:

$$\ell(h^Y, \Phi^Y \mid p_1, \ldots, p_n) = -\sum_i p_i \log h^Y(\Phi^Y(x_i), w_i)$$
$$-\sum_i (1-p_i)\log(1 - h^Y(\Phi^Y(x_i), w_i)) + \alpha^Y \cdot \mathrm{IPM}_{\mathcal{G}^Y}(\{\Phi^Y(x_i)\}_{i:w_i=0}, \{\Phi^Y(x_i)\}_{i:w_i=1}),$$

where $\mathcal{G}^Y$ is a family of functions $g^Y : \mathcal{R}^Y \to \mathcal{Y}$, and α^Y is a hyper-parameter. For two probability density functions p, q defined over $\mathcal{S} \subseteq \mathbb{R}^d$, and for a function family G of functions $g : \mathcal{S} \to \mathbb{R}$, the IPM distance is $\mathrm{IPM}_{\mathrm{G}}(p, q) := \sup_{g \in \mathrm{G}} \left| \int_{\mathcal{S}} g(s)(p(s) - q(s)) ds \right|$. Similarly, we train the potential

Table 2: Performance comparison (MSE ± SD) on synthetic datasets with varying b_D.

	Toy ($b_D = 0$)		Toy ($b_D = 0.5$)		Toy ($b_D = 1$)	
Method	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}
T-learner	0.535 ± 0.041	0.069 ± 0.024	0.514 ± 0.036	0.028 ± 0.017	0.523 ± 0.028	0.109 ± 0.017
CFR	0.536 ± 0.042	0.071 ± 0.025	0.517 ± 0.037	0.025 ± 0.016	0.523 ± 0.028	0.108 ± 0.016
SITE	0.630 ± 0.058	0.023 ± 0.041	0.646 ± 0.077	0.026 ± 0.020	0.654 ± 0.039	0.128 ± 0.045
Dragonnet	0.612 ± 0.080	0.101 ± 0.055	0.499 ± 0.023	0.028 ± 0.024	0.504 ± 0.018	0.095 ± 0.032
CFR-ISW	0.552 ± 0.057	0.064 ± 0.040	0.602 ± 0.084	0.034 ± 0.024	0.590 ± 0.081	0.122 ± 0.023
DR-CFR	0.539 ± 0.030	0.071 ± 0.032	0.521 ± 0.044	0.032 ± 0.026	0.524 ± 0.038	0.107 ± 0.035
DER-CFR	0.548 ± 0.051	0.051 ± 0.029	0.540 ± 0.037	0.066 ± 0.043	0.568 ± 0.034	0.162 ± 0.032
CEVAE	0.661 ± 0.077	0.123 ± 0.039	0.661 ± 0.077	0.122 ± 0.039	0.661 ± 0.077	0.122 ± 0.039
GANITE	0.672 ± 0.074	0.173 ± 0.037	0.662 ± 0.075	0.147 ± 0.036	0.655 ± 0.076	0.122 ± 0.035
T-DF	0.416 ± 0.019	0.021 ± 0.008	0.432 ± 0.013	0.017 ± 0.014	0.407 ± 0.016	0.013 ± 0.007
CFR-DF	**0.409 ± 0.018**	**0.019 ± 0.008**	**0.404 ± 0.014**	**0.013 ± 0.009**	**0.395 ± 0.013**	**0.011 ± 0.009**

Table 3: ϵ_{PEHE} of HTE estimations for potential response times with varying b_D.

Toy ($b_D = 0$)	$\mathbb{P}(D(1) > d \mid Y(0)=1, Y(1)=1, X=x) - \mathbb{P}(D(0) > d \mid Y(0)=1, Y(1)=1, X=x)$						$\tau_D(x)$
$D > d$	$d = 0.1$	$d = 0.2$	$d = 0.5$	$d = 1.0$	$d = 2.0$	$d = 5.0$	N/A
T-DF	0.017 ± 0.003	0.031 ± 0.005	0.056 ± 0.009	0.068 ± 0.012	0.055 ± 0.012	0.015 ± 0.007	0.190 ± 0.030
CFR-DF	**0.014 ± 0.001**	**0.025 ± 0.003**	**0.045 ± 0.005**	**0.054 ± 0.007**	**0.042 ± 0.005**	**0.008 ± 0.002**	**0.152 ± 0.016**
Toy ($b_D = 1$)	$\mathbb{P}(D(1) > d \mid Y(0)=1, Y(1)=1, X=x) - \mathbb{P}(D(0) > d \mid Y(0)=1, Y(1)=1, X=x)$						$\tau_D(x)$
$D > d$	$d = 0.1$	$d = 0.2$	$d = 0.5$	$d = 1.0$	$d = 2.0$	$d = 5.0$	N/A
T-DF	0.025 ± 0.004	0.040 ± 0.007	0.055 ± 0.010	0.054 ± 0.013	0.041 ± 0.014	0.012 ± 0.007	0.321 ± 0.056
CFR-DF	**0.024 ± 0.003**	**0.037 ± 0.005**	**0.048 ± 0.005**	**0.043 ± 0.006**	**0.030 ± 0.006**	**0.006 ± 0.002**	**0.314 ± 0.047**

response time model using the training loss:

$$\ell(h^D, \Phi^D \mid p_1, \ldots, p_n) = \sum_{i:\tilde{y}_i^t=1} \log h^D(\Phi^D(x_i), w_i, d_i)$$

$$+ \sum_{i:\tilde{y}_i^t=0} p_i \log \int_{t_i}^{\infty} h^D(\Phi^D(x_i), w_i, u)du + \alpha^D \cdot \mathrm{IPM}_{\mathcal{G}^D}(\{\Phi^D(x_i)\}_{i:w_i=0}, \{\Phi^D(x_i)\}_{i:w_i=1}),$$

with $\mathcal{G}^D$ and α^D defined similarly. We summarize the whole algorithm including the detailed backbone and hyper-parameters choosing, as well as provide the pseudo-code in Appendix A.3. In addition, our work can be naturally extended to non-binary treatments with the identifiability results of true HTE in all strata, i.e., $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$. See Appendix A.4 for more details.

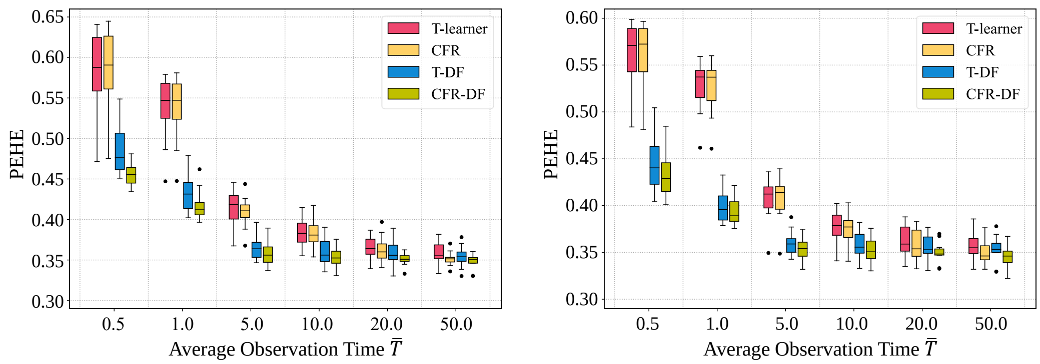
4 Experiments

4.1 Baselines and Evaluation Protocols

We evaluate our framework CFR-DF, and its variant without balancing regularization (T-DF), in the task of (i) estimating HTE on the eventual outcome and (ii) estimating HTE on the response time in the always-positive stratum. We compare our method with the following methods: **T-learner** (Künzel et al., 2019), representation-based algorithms including **CFR** (Shalit et al., 2017), **SITE** (Yao et al., 2018), **Dragonnet** (Shi et al., 2019), **CFR-ISW** (Hassanpour & Greiner, 2019), **DR-CFR** (Hassanpour & Greiner, 2020) and **DER-CFR** (Wu et al., 2022), and generative algorithms **CEVAE** (Louizos et al., 2017) and **GANITE** (Yoon et al., 2018). Following previous studies (Shalit et al., 2017; Wu et al., 2022), we evaluate the performance of HTE estimation using $\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\left((\hat{y}_i(1) - \hat{y}_i(0)) - (y_i(1) - y_i(0))\right)^2$ and $\epsilon_{\text{ATE}} = |\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i(1) - \hat{y}_i(0) - (y_i(1) - y_i(0)))|$, where $\hat{y}_i$ and y_i are predicted and true outcomes.

4.2 Datasets

Synthetic Datasets. Since the true potential outcomes are rarely available for real-world, we conduct simulation studies using synthetic datasets as follows. The observed covariates are generated from $X \sim \mathcal{N}(0, I_{m_X})$, where I_{m_X} denotes m_X-degree identity matrix. The observed treatment



(a) TOY ($b_D = 0$). (b) TOY ($b_D = 1$).

Figure 3: Effects of varying average observation time on synthetic datasets with varying b_D.

$W \sim \text{Bern}(\pi(X))$, where $\pi(X) = \mathbb{P}(W = 1 \mid X) = \sigma(\theta_W \cdot X)$, $\theta_W \sim U(-1, 1)$, and $\sigma(\cdot)$ denotes the sigmoid function. For the eventual potential outcomes, we generate the control outcome $Y(0) \sim \text{Bern}(\sigma(\theta_{Y0} \cdot X^2 + 1))$, and the treated outcome $Y(1) \sim \text{Bern}(\sigma(\theta_{Y1} \cdot X^2 + 2))$, where $\theta_{Y0}, \theta_{Y1} \sim U(-1, 1)$. In addition, we generate the potential response time $D(0) \sim \text{Exp}(\exp(\theta_{D0} \cdot X)^{-1})$, and $D(1) \sim \text{Exp}(\exp(\theta_{D1} \cdot X - b_D)^{-1})$, where $\theta_{D0}, \theta_{D1} \sim U(-0.1, 0.1)$, and b_D *controls the heterogeneity of response time functions*. The observation time is generated via $T \sim \text{Exp}(\lambda)$, where λ is the rate parameter of the exponential distribution, and we set $\lambda = 1$ in our experiments, i.e., the average observation time is $\bar{T} = \lambda^{-1} = 1$. Finally, the observed outcome is $\tilde{Y}^T(W) = W \cdot Y(1) \cdot \mathbb{I}(T \geq D(1)) + (1 - W) \cdot Y(0) \cdot \mathbb{I}(T \geq D(0))$, where $\mathbb{I}(\cdot)$ is the indicator function. Based on the data generation process described above, we sample $N = 20,000$ samples for training and $3,000$ samples for testing. We repeat each experiment 10 times to report the mean and standard deviation of the results (ϵ_{PEHE} and ϵ_{ATE}). Moreover, we vary the heterogeneity of response times by setting $b_D \in \{0, 0.5, 1\}$, named the dataset as TOY ($b_D = 0$), TOY ($b_D = 0.5$), and TOY ($b_D = 1$), respectively. Besides, we evaluate our algorithm on the TOY ($b_D = 0$) and TOY ($b_D = 1$) with the average observation time $\bar{T} \in \{0.5, 1, 5, 10, 20, 50\}$.

Real-World Datasets. We also evaluate our CFR-DF on three widely-adopted real-world datasets: AIDS[1] (Hammer et al., 1997; Norcliffe et al., 2023), JOBS[2] (LaLonde, 1986; Shalit et al., 2017), and TWINS[3] (Almond et al., 2005; Wu et al., 2022). The AIDS dataset collected between January 1996 and January 1997 involved 1,156 patients in 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the United States and Puerto Rico and was used to study the impact and effectiveness of antiretroviral therapy on HIV-positive patients. The JOBS dataset is widely used in the field of causal inference. It is built upon randomized controlled trials and aims to assess the effects of job training programs on employment status. The TWINS dataset is derived from all twins born in the USA between the years 1989 and 1991 and is utilized to assess the influence of birth weight on mortality within one year, from which we obtain covariates X. Following the same procedure for generating synthetic datasets, we generate treatment W, potential outcomes $Y(0)$ and $Y(1)$, potential response times $D(0)$ and $D(1)$, observation time T and factual outcomes $\tilde{Y}^T(W)$. Then we randomly split the samples into training/testing with an 80/20 ratio with 10 repetitions.

4.3 RESULTS

Performance Comparison. We compare our method with the baselines for estimating the HTE on the eventual outcome with varying response time functions in Table 2. The optimal and second-optimal performance are **bold** and underlined, respectively. First, the proposed CFR-DF stably outperforms the baselines, as the previous methods do not take into account the delayed response, leading to biased estimates of HTE. Second, the T-DF method without using balancing regularization slightly degrades the performance compared to CFR-DF, due to the inability to resolve the confounding

[1] https://scikit-survival.readthedocs.io/
[2] http://www.fredjo.com/
[3] http://www.nber.org/data/

Table 4: Performance comparison (MSE ± SD) on JOBS and TWINS datasets.

	AIDS		JOBS		TWINS	
Method	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}
T-learner	0.525 ± 0.052	0.091 ± 0.064	0.528 ± 0.043	0.085 ± 0.041	0.390 ± 0.071	0.050 ± 0.029
CFR	0.531 ± 0.046	0.083 ± 0.058	0.510 ± 0.035	0.064 ± 0.039	0.378 ± 0.057	<u>0.029 ± 0.018</u>
SITE	0.601 ± 0.031	0.082 ± 0.056	0.568 ± 0.045	0.064 ± 0.053	0.495 ± 0.087	0.139 ± 0.053
Dragonnet	0.546 ± 0.051	0.105 ± 0.042	0.555 ± 0.060	0.084 ± 0.060	0.440 ± 0.103	0.096 ± 0.067
CFR-ISW	0.592 ± 0.053	0.098 ± 0.032	0.499 ± 0.035	0.058 ± 0.056	0.392 ± 0.048	0.039 ± 0.023
DR-CFR	0.577 ± 0.056	0.078 ± 0.044	0.525 ± 0.077	0.079 ± 0.060	0.390 ± 0.046	0.039 ± 0.027
DER-CFR	0.609 ± 0.076	0.081 ± 0.074	0.503 ± 0.037	0.072 ± 0.043	0.398 ± 0.068	0.080 ± 0.066
CEVAE	0.623 ± 0.042	0.143 ± 0.019	0.638 ± 0.062	0.102 ± 0.058	0.526 ± 0.055	0.139 ± 0.027
GANITE	0.605 ± 0.034	0.136 ± 0.020	0.629 ± 0.053	0.151 ± 0.067	0.509 ± 0.056	0.139 ± 0.040
T-DF	<u>0.521 ± 0.042</u>	<u>0.077 ± 0.030</u>	<u>0.453 ± 0.066</u>	<u>0.058 ± 0.030</u>	<u>0.366 ± 0.027</u>	0.030 ± 0.018
CFR-DF	**0.499 ± 0.055**	**0.073 ± 0.031**	**0.438 ± 0.059**	**0.051 ± 0.031**	**0.357 ± 0.017**	**0.027 ± 0.015**

bias from covariate shift. Third, we observe a decrease in ϵ_{PEHE} and ϵ_{ATE} of 23% and 17% in TOY ($b_D = 0$), 21% and 48% in TOY ($b_D = 0.5$), and 46% and 88% in TOY ($b_D = 1$), respectively, when comparing our CFR-DF method to the optimal baseline method. These results highlight the scalability of our method to different levels of observation times, demonstrating its potential for real-world applications. Table 3 shows the performance of our methods in estimating HTE on the response times, as described in Section 2.2. We report the ϵ_{PEHE} on estimating $\mathbb{P}(D(1) > d \mid Y(0) = 1, Y(1) = 1, X = x) - \mathbb{P}(D(0) > d \mid Y(0) = 1, Y(1) = 1, X = x)$ and $\tau_D(x)$, respectively, where the former has a more fine-grained description with varying d. We find both T-DF and CFR-DF can effectively estimate the treatment effect on response time, and CFR-DF with balancing regularization stably performs better, again demonstrating the need to adjust for confounding bias. See Appendix A.5.2 for more experiment results with various number of features.

Ablation Studies. Figure 3 compares the proposed CFR-DF and its ablated versions for estimating HTE on the eventual outcome with varying average observation time, where T-DF does not perform balancing regularization, CFR does not consider delayed response, and neither is considered for T-learner. We have the following findings. The proposed CFR-DF and T-DF have significantly better performance when the observation time is shorter, due to their effective adjustment for delayed response. When increasing the average observation time leads to more delayed responses being observed, we find improved performance for all four methods. The ϵ_{PEHE} of CFR-DF stabilizes when the average observation time is above 5, and the variance gradually decreases with increasing observation time. When the observation time reaches 50, meaning all delayed responses have been observed, our method performs similarly to the CFR algorithm, and T-DF is degenerate to T-learner.

Real-World Experiments. We conduct real-world experiments using AIDS, JOBS and TWINS datasets. The AIDS (Hammer et al., 1997) contains people with HIV and SEER with Prostate Cancer. The JOBS dataset (LaLonde, 1986) is based on the National Supported Work program and examines the effects of job training on income and employment status after training. The TWINS dataset (Almond et al., 2005) studies the effects of infant weight on the death rate. Notably, job training takes time to cause changes in incomes, and infants also take time to observe their mortality outcomes (and thus study the effect on mortality), therefore it is reasonable to study the delayed response in such real-world applications. Table 4 demonstrates that CFR-DF outperforms all baselines on these real-world datasets, showcasing its effectiveness.

5 CONCLUSION

This paper studies the HTE estimation problem by further considering the response time needed for a treatment to produce a causal effect on the outcome. Specifically, we propose a principled learning algorithm, called CFR-DF, to estimate both eventual potential outcomes and potential response times. Considering the widespread delayed feedback outcomes, we believe such a study is meaningful for real-world applications. A shortcoming of our study is the validity of the assumptions in practice, e.g., we need enough observation time to identify HTE on the eventual potential outcome, and principal ignorability is further required to identify HTE on the response time. Studying how to weaken these assumptions, and identifying and estimating HTE with delayed responses are served as future topics.

REFERENCES

Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

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

Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric utilities. *arXiv preprint arXiv:2206.10479*, 2022.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1105, 2014.

Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145, 2021.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Alicia Curth and Mihaela van der Schaar. Understanding the impact of competing events on heterogeneous treatment effect estimation from time-to-event data. In *International Conference on Artificial Intelligence and Statistics*, pp. 7961–7980. PMLR, 2023.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34:26740–26753, 2021.

Constantine Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Muskan Gupta, Gokul Kannan, Ranjitha Prasad, and Garima Gupta. Deep survival analysis and counterfactual inference using balanced representations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Scott M Hammer, Kathleen E Squires, Michael D Hughes, Janet M Grimes, Lisa M Demeter, Judith S Currier, Joseph J Eron Jr, Judith E Feinberg, Henry H Balfour Jr, Lawrence R Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

Alexis Hannart, J Pearl, FEL Otto, P Naveau, and M Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99–110, 2016.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HkxBJT4YvB`.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81: 945–960, 1986.

Kosuke Imai and Zhichao Jiang. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*, 2020.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46, pp. 79–95, 2012.

Zhichao Jiang, Shu Yang, and Peng Ding. Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3768–3774, 2016.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, second edition, 2015.

Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. Counterfactual phenotyping with censored time-to-events. *arXiv preprint arXiv:2202.11089*, 2022.

Chirag Nagpal, Vedant Sanil, and Artur Dubrawski. Recovering sparse and interpretable subgroups with heterogeneous treatment effects with censored time-to-event outcomes. *Proceedings of Machine Learning Research vol TBD*, 1:18, 2023.

Jerzy Splawa Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472, 1990.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

Alexander Norcliffe, Bogdan Cebere, Fergus Imrie, Pietro Lio, and Mihaela van der Schaar. Survivalgan: Generating time-to-event data for survival analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 10279–10304. PMLR, 2023.

Judea Pearl. Principal stratification – a goal or a tool? *The International Journal of Biostatistics*, 7 (1):1–13, 2011.

Paul R Rosenbaum. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

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

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66:688–701, 1974.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *ICML*, 2016.

Stefan Schrod, Andreas Schäfer, Stefan Solbrig, Robert Lohmayer, Wolfram Gronwald, Peter J Oefner, Tim Beißbarth, Rainer Spang, Helena U Zacharias, and Michael Altenbuchinger. Bites: balanced individual treatment effect for survival data. *Bioinformatics*, 38(Supplement_1):i60–i67, 2022.

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4989–5001, 2022.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Yuya Yoshikawa and Yusaku Imai. A nonparametric delayed feedback model for conversion rate prediction. *arXiv preprint arXiv:1802.00255*, 2018.

A Appendix

A.1 Related Work

In Heterogeneous treatment effect (HTE) estimation, non-random treatment assignments can result in different probabilities of missing covariates in different treatment arms, which may introduce confounding bias. To address this issue, most methods strive to balance covariates to estimate HTE accurately, such as matching, stratification, outcome regression, weighting, and doubly robust methods (Rosenbaum, 1987; Rosenbaum & Rubin, 1983; Li et al., 2016; Hainmueller, 2012). With the advances in deep learning, Balancing Neural Network (BNN) (Johansson et al., 2016) and CounterFactual Regression (CFR) (Shalit et al., 2017) propose to learn a covariate representation that is independent of the treatment to overcome the covariate shift between the treatment and control groups, in which the independence is measured by Integral Probability Metric (IPM) (Johansson et al., 2016; Shalit et al., 2017). SITE (Yao et al., 2018) preserves local similarity and balances the distributions of the representation simultaneously. Motivated by targeted learning (van der Laan & Rose, 2011), DragonNet (Shi et al., 2019) proposed an adaptive neural network to end-to-end model propensity scores and counterfactual outcomes. DR-CFR (Hassanpour & Greiner, 2020) and DeR-CFR (Wu et al., 2022) propose a disentanglement framework to identify the representation of confounders from all observed variables. By exploiting the generative models, CEVAE (Louizos et al., 2017) and GANITE (Yoon et al., 2018) generate counterfactual outcomes for HTE estimation. However, these algorithms rely on timely and accurate observation of the eventual potential outcomes.

In practice, interventions usually take time to have a causal effect on the outcome (Chapelle, 2014; Yoshikawa & Imai, 2018). Despite the problem setup and the causal estimand of interest is different, many studies have examined HTE estimation under time-to-event data. Curth et al. (2021) used neural networks for discrete time analyses and Chapfuwa et al. (2021) used generative models for counterfactual time-to-event data analysis in continuous time. Based on the Cox model (Cox, 1972), Schrod et al. (2022) proposed a treatment-specific semi-parametric Cox loss using time-to-event data for treatment optimization. Gupta et al. (2023) derived a binary treatment evidence lower bound (ELBO) for parametric survival analysis, and designed a neural network for learning the per-individual survival density. Different from Chapfuwa et al. (2021); Curth et al. (2021), Curth & van der Schaar (2023) considered time-to-event data with competing events, which can act as an additional source of covariate shift. In addition, Nagpal et al. (2022) presented a latent variable approach to mediate the base survival rates and help determine the effects of an intervention. Nagpal et al. (2023) extended Nagpal et al. (2022) by proposing a statistical approach to recovering sparse phenogroups (or subtypes) that demonstrate differential treatment effects as compared to the study population. Though delayed response can be considered as a right-censored problem, rather than focusing on the effect of treatment on survival curves, this paper assumes that it takes time to yield an observable outcome that eventually has a positive outcome (e.g., conversion in uplift modeling) and considers both conversion time and whether or not to convert as potential outcomes by utilizing a *hybrid model*. By considering the joint potential outcome of individuals from a principal stratification perspective (Frangakis & Rubin, 2002; Pearl, 2011), we theoretically prove that the potential response times on subgroups in which individuals always have positive eventual outcomes regardless of treatment are identifiable. It is also interesting to note that the problem studied in this paper can also be considered as a noisy label on the eventual outcome of interest due to the limited observation time, which causes the previous HTE methods to be biased.

A.2 Theorems and Proofs

A.2.1 The Proofs of Theorems 1 and 2

First, we recap the assumptions in Section 2.3 as below. Next, we provide formal proofs of Theorem 1, Lemma 1, and Theorem 2, respectively.

Assumption 1 (Unconfoundedness)**.** There is no unmeasured confounders, $W \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1)) \mid X$ *for all* $t > 0$.

Assumption 2 (Time Independence)**.** Time T is independent of potentials, $T \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1), W) \mid X$ *for all* $t > 0$.

Assumption 3 (Time Sufficiency)**.** $\inf\{d : F_D^{(w)}(d \mid Y(w)=1, X)=1\} < \inf\{t : F_T(t)=1\}$ *for* $w=0,1$*, where* $F(\cdot)$ *is the cumulative distribution function (cdf).*

Assumption 4 (Monotonicity)**.** $Y(0) \leq Y(1)$.

Assumption 5 (Principal Ignorability)**.** $(W, Y(w)) \perp\!\!\!\perp D(1-w) \mid Y(1-w), X$ *for* $w=0,1$.

Theorem 1. *Under Assumptions 1-3, the HTE on the eventual outcome* $\tau(x)$ *is identifiable.*

Proof of Theorem 1. For units with $Y(w)=0$, we set $D(w)=\infty$, for $w=0,1$. We first prove the identifiability of $\mathbb{P}(D(w)>t \mid X=x)$ for $w=0,1$ and $t>0$. Under Assumption 1, we have:

$$\begin{aligned}
&-\frac{d}{dt}\log \mathbb{P}(D(w)>t \mid X=x) && (1)\\
&=\lim_{h\to 0^+}\frac{\frac{1}{h}\mathbb{P}(t<D(w)\leq t+h \mid X=x)}{P(D(w)>t \mid X=x)}\\
&=\lim_{h\to 0^+}\frac{\frac{1}{h}\mathbb{P}(t<D(w)\leq t+h \mid W=w, X=x)}{\mathbb{P}(D(w)>t \mid W=w, X=x)} && (2)\\
&=\lim_{h\to 0^+}\frac{1}{h}\mathbb{P}(t<D(w)\leq t+h \mid W=w, X=x, D(w)>t),
\end{aligned}$$

where the first equality follows from the definition of first-order derivative, the second equality follows from the unconfoundedness assumption, and the third equality follows from the definition of conditional probability. Under Assumption 2, we obtain the identifiability result in the following:

$$\begin{aligned}
&\lim_{h\to 0^+}\frac{1}{h}\mathbb{P}(t<D(w)\leq t+h \mid W=w, X=x, D(w)>t)\\
&=\lim_{h\to 0^+}\frac{1}{h}\mathbb{P}(t<D(w)\leq t+h \mid W=w, X=x, D(w)>t, T>t)\\
&=\lim_{h\to 0^+}\frac{1}{h}\mathbb{P}(t<\min\{D(w),T\}\leq t+h, \mathbb{I}(D(w)\leq T)=1 \mid \text{cond})\\
&=\lim_{h\to 0^+}\frac{1}{h}\mathbb{P}(t<\min\{D,T\}\leq t+h, \mathbb{I}(D\leq T)=1 \mid \text{cond}), && (3)
\end{aligned}$$

where cond $=\{W=w, X=x, \min\{D,T\}>t\}$, and the first equality follows from the time independence assumption, the second equality follows from the equivalence between $t<D(w)\leq t+h$ and $t<\min\{D(w),T\}\leq t+h$ and $D(w)\leq T$, given the condition that $T>t$ with a sufficiently small time period $h\to 0^+$, the third equality follows from the unconfoundedness assumption. Also, we can identify:

$$\mathbb{P}(D(w)>t \mid X=x)=\exp\left\{\int_0^t \frac{d}{du}\log \mathbb{P}(D(w)>u \mid X=x)du\right\} \tag{4}$$

for $w=0,1$, because we have $-\frac{d}{dt}\log \mathbb{P}(D(w)>t \mid X=x)$.

We next show the identifiability of $\mathbb{P}(Y(w)=1 \mid X=x)$. Under Assumption 3, we have

$$\begin{aligned}
&\mathbb{P}(Y(w)=1 \mid X=x)=1-\mathbb{P}(Y(w)=0 \mid X=x)\\
&=1-\lim_{t\to\infty}\mathbb{P}(D(w)>t \mid X=x)\\
&=1-\mathbb{P}(D(w)>q_d \mid X=x)=1-\mathbb{P}(D(w)>q \mid X=x) && (5)
\end{aligned}$$

for $q_d \leq q < q_t$, where $q_d=\inf\left\{d : F_D^{(w)}(d \mid Y(w)=1, X)=1\right\}$, $q_t=\inf\{t : F_T(t)=1\}$ and $F(\cdot)$ is the cumulative distribution function (cdf). Therefore, $\mathbb{P}(Y(w)=1 \mid X=x)$ is identifiable from observed data for $w=0,1$. $\square$

Lemma 1. *Under Assumptions 1-4,* $\mathbb{P}(Y(0)=1, Y(1)=1 \mid X=x)$ *is identifiable.*

Proof of Lemma 1. Under Assumption 4, we have

$$\begin{aligned}
&\mathbb{P}(Y(0)=0, Y(1)=0 \mid X=x) = \mathbb{P}(Y(1)=0 \mid X=x) \\
&\mathbb{P}(Y(0)=0, Y(1)=1 \mid X=x) = \mathbb{P}(Y(1)=1 \mid X=x) - \mathbb{P}(Y(0)=1 \mid X=x) \\
&\mathbb{P}(Y(0)=1, Y(1)=1 \mid X=x) = \mathbb{P}(Y(0)=1 \mid X=x).
\end{aligned} \tag{6}$$

Then the identifiability of the left-hand side parameters follows directly from the identifiability of $\mathbb{P}(Y(w)=1 \mid X=x)$ for $w=0,1$ under Assumptions 1-3 as shown in Theorem 1. □

Theorem 2. *Under Assumptions 1-5, the HTE on the response time in the always-positive stratum* $\tau_D(x) = \mathbb{E}[D(1)-D(0) \mid Y(0)=1, Y(1)=1, X=x]$ *is identifiable.*

Proof of Theorem 2. Under Assumption 5, i.e., $(W, Y(0)) \perp\!\!\!\perp D(1) \mid Y(1), X$, we have

$$\begin{aligned}
&\mathbb{P}(D(1)<t \mid Y(0)=1, Y(1)=1, X=x) = \mathbb{P}(D(1)<t \mid Y(1)=1, X=x) \\
&= \mathbb{P}(D(1)<t \mid Y(1)=1, X=x, W=1) = \mathbb{P}(D(1)<t \mid Y=1, X=x, W=1) \\
&= \frac{\mathbb{P}(D<t \mid X=x, W=1)}{\mathbb{P}(Y=1 \mid X=x, W=1)} \\
&= \frac{1-\exp\left\{\int_0^t \frac{d}{du}\log\mathbb{P}(D(1)>u \mid X=x)du\right\}}{1-\lim_{t\to\infty}\exp\left\{\int_0^t \frac{d}{du}\log\mathbb{P}(D(1)>u \mid X=x)du\right\}},
\end{aligned} \tag{7}$$

which is identifiable, because we have proved the identifiability of $-\frac{d}{dt}\log\mathbb{P}(D(1)>t \mid X=x)$ in Theorem 1. Similarly, we can identify

$$\begin{aligned}
&\mathbb{P}(D(0)<t \mid Y(0)=1, Y(1)=1, X=x) = \\
&\frac{1-\exp\left\{\int_0^t \frac{d}{du}\log\mathbb{P}(D(0)>u \mid X=x)du\right\}}{1-\lim_{t\to\infty}\exp\left\{\int_0^t \frac{d}{du}\log\mathbb{P}(D(0)>u \mid X=x)du\right\}}.
\end{aligned} \tag{8}$$

Then $\tau_D(x)$ is identifiable due to

$$\begin{aligned}
\tau_D(x) =& \mathbb{E}[D(1)-D(0) \mid Y(0)=1, Y(1)=1, X=x] \\
=& -\int_0^\infty \mathbb{P}(D(1)<u \mid Y(0)=1, Y(1)=1, X=x)du \\
&+ \int_0^\infty \mathbb{P}(D(0)<u \mid Y(0)=1, Y(1)=1, X=x)du.
\end{aligned} \tag{9}$$

□

A.2.2 Computation of (Non-)Parametric Potential Response Time Models

In this paper, we propose a principled learning approach called CFR-DF (**C**ounter**F**actual **R**egression with **D**elayed **F**eedback) that simultaneously predicts potential outcomes and potential response times by employing an EM algorithm with eventual outcomes treated as latent variables. Due to space limitations, we only provide the explicit solutions of the EM algorithm in a general functional form for estimating the parameters of interest in Section 3 in the main text. However, in practice, empirical computation requires model specification: either (i) a parametric model or (ii) a non-parametric model based on weighted kernel functions.

Parametric model: One can assume that the potential delayed response times obey exponential models for both treatment and control groups. Specifically, let $\mathbb{P}(D(w)=u \mid X=\mathbf{x}, Y(w)=1) = \lambda_w(\mathbf{x})\exp(-\lambda_w(\mathbf{x})u)$ for $w=0,1$. Then we have:

$$\begin{aligned}
&\int_t^\infty \mathbb{P}(D(w)=u \mid X=\mathbf{x}, Y(w)=1)du \\
&= \int_t^\infty \lambda_w(\mathbf{x})\exp(-\lambda_w(\mathbf{x})u)\,du = \exp(-\lambda_w(\mathbf{x})t)
\end{aligned} \tag{10}$$

in the derived p_i in the E-step.

Non-parametric model based on weighted kernel functions: potential delayed response times can be further extended to a nonparametric model using a set of weighted kernel functions. Specifically, let the non-parametric hazard function is $h_w(d;\mathbf{x}) = \sum_{l=1}^{L} \alpha_l^w(\mathbf{x}) k\left(t_l, d\right) \quad \text{for} \quad w=0,1$, where k is a kernel function returning a positive value, and intuitively represents the similarity between two time points. Here, one can use kernel functions as k such that $k\left(t_l, u\right), \int_0^a k\left(t_l, u\right) du$ and $\int_a^\infty k\left(t_l, u\right) du$ for $t_l, u, a \geq 0$ can be calculated analytically.

For example, a Gaussian kernel with bandwidth parameter $h > 0$ leads to

$$k\left(t_l, u\right) = \exp\left(-\frac{\left(t_l - u\right)^2}{2h^2}\right), \tag{11}$$

$$\int_0^a k\left(t_l, u\right) du = -h\sqrt{\frac{\pi}{2}}\left[\operatorname{erf}\left(\frac{t_l - a}{\sqrt{2}h}\right) - \operatorname{erf}\left(\frac{t_l}{\sqrt{2}h}\right)\right] \tag{12}$$

$$\int_a^\infty k\left(t_l, u\right) du = h\sqrt{\frac{\pi}{2}}\left[1 + \operatorname{erf}\left(\frac{t_l - a}{\sqrt{2}h}\right)\right], \tag{13}$$

where leads to the analytical form p_i in the E-step.

Given the hidden variable values p_i computed from the E-step, we can plug them into the expected log-likelihood during the M-step:

$$\begin{aligned}
&\sum_{i:\tilde{y}_i^t=1} \log \mathbb{P}(\tilde{Y}_i^T = 1, D = d_i \mid X = x_i, W = w_i, T = t_i) \\
+ &\sum_{i:\tilde{y}_i^t=0} (1 - p_i) \log \mathbb{P}(\tilde{Y}_i^T = 0, Y_i(w_i) = 0 \mid X = x_i, W = w_i, T = t_i) \\
+ &\sum_{i:\tilde{y}_i^t=0} p_i \log \mathbb{P}(\tilde{Y}_i^T = 0, Y_i(w_i) = 1 \mid X = x_i, W = w_i, T = t_i).
\end{aligned} \tag{14}$$

From a similar argument as derived above, the expected log-likelihood is equal to:

$$\begin{aligned}
&\sum_i p_i \log \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i) + (1 - p_i) \log(1 - \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i)) \\
+ &\sum_{i:\tilde{y}_i^t=1} \log \mathbb{P}(D_i(w_i) = d_i \mid X = x_i, Y_i(w_i) = 1) \\
+ &\sum_{i:\tilde{y}_i^t=0} p_i \log \int_{t_i}^\infty \mathbb{P}(D(w_i) = u \mid X = x_i, Y_i(w_i) = 1) du,
\end{aligned} \tag{15}$$

in which the eventual potential outcome model $\mathbb{P}(Y(w) = 1 \mid X = x)$ and the potential response time model $\mathbb{P}(D(w) = d \mid X = x, Y(w) = 1)$ can be optimized independently. In our experiments, we used *Parametric models* for delay time modeling in the treated and control groups.

A.3 Algorithm, Hyper-Parameters and Discussion

A.3.1 Algorithm Details and Environment Configuration

Motivation: In this paper, we study the problem of estimating HTE with a delayed response, which can be seen as a censoring problem with imbalanced treatment assignment: the observation time T refers to the "time-to-censor", the response time D refers to the "time-to-event", and the treatment is not assigned at random. We must emphasize that simply applying the expectation-maximization technique is insufficient to recover the delayed outcome without making additional assumptions and identification guarantees. Because this problem involves not only missing data but also survival analysis and confounding bias. To address these issues, we propose a novel CFR-DF approach that extends counterfactual regression to delayed feedback outcomes using a modified EM algorithm with identification guarantees. In Appendix A.2.2, we provide the explicit solutions of the EM algorithm with model specification: either (i) a parametric model or (ii) a non-parametric model based on weighted kernel functions. In our experiments, we use *Parametric models* for delay time modeling in the treated and control groups. Algorithm 1 shows the pseudo-code of our CFR-DF.

Implementation of CFR-DF. In the CFR-DF architecture (Figure 4), we use three-layer neural networks Φ_0^Y and Φ_1^Y with ELU activation function and BatchNorm to learn representation of the eventual outcome, and two-layer neural networks Φ_0^D and Φ_1^D with ELU activation function and BatchNorm to learn representation of the delayed response time. Each layer in these networks consists of m_X neural units. Then, we use a single-layer network h^Y with Sigmoid activation to achieve $\hat{P}(Y=1)$ and a single-layer network h^D with SoftPlus sigmoid activation to achieve $\hat{\lambda}$. Dropout is not utilized in the CFR-DF architecture, but BatchNorm is applied in each layer of the representation networks. Finally, we update $\left\{\Phi_0^Y, \Phi_1^Y, \Phi_0^D, \Phi_1^D, h^D, h^Y\right\}$ using Adam L_s optimizer.

Based on the developed EM algorithm in a general functional form for estimating the parameters of interest, we now show the empirical computation details for both (i) parametric model and (ii) non-parametric model based on weighted kernel functions.

• **Parametric model**: One can assume that the potential delayed response times obey exponential models for both treatment and control groups. Specifically, let $\mathbb{P}(D(w)=u \mid X=\mathbf{x}, Y(w)=1)=\lambda_w(\mathbf{x}) \exp\left(-\lambda_w(\mathbf{x})u\right)$ for $w=0,1$, we have $\int_t^{\infty} \mathbb{P}(D(w)=u \mid X=\mathbf{x}, Y(w)=1)du = \int_t^{\infty} \lambda_w(\mathbf{x}) \exp\left(-\lambda_w(\mathbf{x})u\right) du = \exp\left(-\lambda_w(\mathbf{x})t\right)$ in the derived p_i in the E-step.

• **Non-parametric model based on weighted kernel functions**: The estimation of potential delayed response times can be further extended to a nonparametric model using a set of weighted kernel functions. Specifically, let the non-parametric hazard function is $h_w(d;\mathbf{x}) = \sum_{l=1}^{L} \alpha_l^w(\mathbf{x}) k\left(t_l, d\right)$ for $w=0,1$, where k is a kernel function returning a positive value, and intuitively represents the similarity between two time points. Here, one can use kernel functions as k such that $k\left(t_l, u\right), \int_0^a k\left(t_l, u\right) du$ and $\int_a^{\infty} k\left(t_l, u\right) du$ for $t_l, u, a \geq 0$ can be calculated analytically. For example, a Gaussian kernel with bandwidth parameter $h>0$ leads to $k\left(t_l, u\right) = \exp\left(-\frac{(t_l-u)^2}{2h^2}\right)$, $\int_0^a k\left(t_l, u\right) du = -h\sqrt{\frac{\pi}{2}}\left[\operatorname{erf}\left(\frac{t_l-a}{\sqrt{2}h}\right) - \operatorname{erf}\left(\frac{t_l}{\sqrt{2}h}\right)\right]$, and $\int_a^{\infty} k\left(t_l, u\right) du = h\sqrt{\frac{\pi}{2}}\left[1+\operatorname{erf}\left(\frac{t_l-a}{\sqrt{2}h}\right)\right]$, where leads to the analytical form p_i in the E-step.

Given the hidden variable values p_i computed from the E-step, we can plug them into the expected log-likelihood at the M-step:

$$\begin{aligned}
&\sum_{i:\tilde{y}_i^t=1} \log \mathbb{P}(\tilde{Y}_i^T=1, D=d_i \mid X=x_i, W=w_i, T=t_i) \\
&+\sum_{i:\tilde{y}_i^t=0} (1-p_i) \log \mathbb{P}(\tilde{Y}_i^T=0, Y_i(w_i)=0 \mid X=x_i, W=w_i, T=t_i) \\
&+\sum_{i:\tilde{y}_i^t=0} p_i \log \mathbb{P}(\tilde{Y}_i^T=0, Y_i(w_i)=1 \mid X=x_i, W=w_i, T=t_i).
\end{aligned}$$

From a similar argument as derived above, the expected log-likelihood is equal to:

$$\begin{aligned}
&\sum_i p_i \log \mathbb{P}(Y_i(w_i)=1 \mid X=x_i) \\
&+\sum_i (1-p_i) \log(1-\mathbb{P}(Y_i(w_i)=1 \mid X=x_i)) \\
&+\sum_{i:\tilde{y}_i^t=1} \log p(D_i(w_i)=d_i \mid X=x_i, Y_i(w_i)=1) \\
&+\sum_{i:\tilde{y}_i^t=0} p_i \log \int_{t_i}^{\infty} p(D(w_i)=u \mid X=x_i, Y_i(w_i)=1) du,
\end{aligned}$$

in which the eventual potential outcome model $\mathbb{P}(Y(w)=1 \mid X=x)$ and the potential response time model $p(D(w)=d \mid X=x, Y(w)=1)$ can be optimized independently. Notably, in our experiments, we used *Parametric models* for delay time modeling in the treated and control groups.

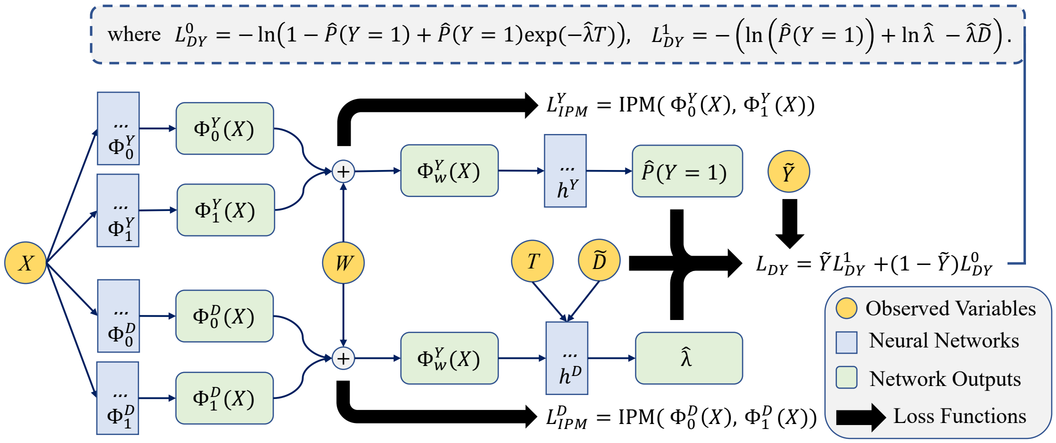


Figure 4: Overview of CFR-DF Architecture. For the representation block, we use multi-layer neural networks Φ with ELU activation function to learn representation and each network has two/three layers with m_X units, respectively. Then, we use a single-layer network h^Y with Sigmoid activation to achieve $\hat{P}(Y=1)$ and a single-layer network h^D with SoftPlus sigmoid activation to achieve $\hat{\lambda}$.

Algorithm 1 CounterFactual Regression with Delayed Feedback Outcomes (CFR-DF)

Input: Observational data $\mathbb{D} = \{x_i, w_i, t_i, \tilde{d}_i, \tilde{y}_i\}_{i=1}^n$ (we set $\tilde{d}_i = -1$ for all subjects with $\tilde{y}_i = 0$ in training process); hyper-parameters α^Y and α^D; neural networks $\{\Phi_0^Y(\cdot), \Phi_1^Y(\cdot), \Phi_0^D(\cdot), \Phi_1^D(\cdot), h^D(\cdot), h^Y(\cdot)\}$; maximum number of iterations $M = 3000$; stopping criterion $\epsilon = 0.002$; initiation loss $L_{s=0} = 9999.9$; and iteration counter $s = 0$.
Output: $\hat{P}_i(Y=1) = h^Y(\Phi^Y(x_i), w_i)$, $\quad \hat{d}_i = \hat{\lambda}_i^{-1}$, $\quad \hat{\lambda}_i = h^D(\Phi^D(x_i), w_i, t_i)$.
Loss function: $L = \tilde{Y} \cdot L_{DY}^1 + (1-\tilde{Y}) \cdot L_{DY}^0 + \alpha^D \cdot L_{IPM}^D + \alpha^Y \cdot L_{IPM}^Y$.
CFR-DF:
$s \leftarrow s+1$;
$L_s = \tilde{Y} \cdot L_{DY}^1 + (1-\tilde{Y}) \cdot L_{DY}^0 + \alpha^D \cdot L_{IPM}^D + \alpha^Y \cdot L_{IPM}^Y$;
while $s \leq M$ and $|L_s - L_{s-1}| > \epsilon$ **do**
  $s \leftarrow s+1$;
  $\Phi^Y(x_i) = w_i\Phi_1^Y(x_i) + (1-w_i)\Phi_0^Y(x_i)$, $\quad \Phi^D(x_i) = w_i\Phi_1^D(x_i) + (1-w_i)\Phi_0^D(x_i)$;
  $\hat{P}_i(Y=1) = h^Y(\Phi^Y(x_i), w_i)$, $\quad \hat{\lambda}_i = h^D(\Phi^D(x_i), w_i, t_i)$;
  $L_{IPM}^Y = \text{IPM}\left(\{\Phi^Y(x_i)\}_{i:w_i=0}, \{\Phi^Y(x_i)\}_{i:w_i=1}\right)$;
  $L_{IPM}^D = \text{IPM}\left(\{\Phi^D(x_i)\}_{i:w_i=0}, \{\Phi^D(x_i)\}_{i:w_i=1}\right)$;
  $L_{DY}^0(x_i) = -\ln(1-\hat{P}_i(Y=1) + \hat{P}_i(Y=1)\exp(-\hat{\lambda}_i t_i))$;
  $L_{DY}^1(x_i) = -(\ln(\hat{P}_i(Y=1)) + \ln\hat{\lambda}_i - \hat{\lambda}_i\tilde{d}_i)$;
  $L_s = \frac{1}{n} \cdot \sum_{i=1}^n \left(\hat{y}_i L_{DY}^1(x_i) + (1-\hat{y}_i)L_{DY}^0(x_i)\right) + \alpha^D \cdot L_{IPM}^D + \alpha^Y \cdot L_{IPM}^Y$;
  Update $\{\Phi_0^Y, \Phi_1^Y, \Phi_0^D, \Phi_1^D, h^D, h^Y\} \leftarrow \text{Adam}\{L_s\}$;
end while

Hardware used: Ubuntu 16.04.3 LTS operating system with 2 * Intel Xeon E5-2660 v3 @ 2.60GHz CPU (40 CPU cores, 10 cores per physical CPU, 2 threads per core), 256 GB of RAM, and 4 * GeForce GTX TITAN X GPU with 12GB of VRAM.

Software used: Python 3.8 with numpy 1.24.2, pandas 2.0.0, pytorch 2.0.0.

A.3.2 Hyper-Parameter Optimization

In this paper, we adopt an early stopping criterion (ε) to select the best-evaluated iterate for each model. The hyper-parameters α^Y and α^D are selected from a range of values $\{1\text{e}-4, 5\text{e}-4, 1\text{e}-3, 5\text{e}-3, 1\text{e}-2, 5\text{e}-2, 1\text{e}-1, 1.00\}$ based on the mean squared error (MSE) of $Y(1)$ on the training data. We optimize the hyper-parameters in CFR-DF by minimizing the objective loss on the training data. Taking Toy($m_X = 20$) as an example, as depicted in Figure 5, we determine the

Table 5: Optimal Hyper-Parameters.

	α^Y	α^D
TOY($m_X = 5$)	0.005	0.1
TOY($m_X = 10$)	0.01	0.1
TOY($m_X = 20$)	0.01	0.1
TOY($m_X = 40$)	0.01	0.1
AIDS	0.01	0.01
JOBS	0.005	0.05
TWINS	0.005	0.01

Table 6: Datasets Used for Evaluation.

	No. instances	No. features
TOY($m_X = 5$)	20000	5
TOY($m_X = 10$)	20000	10
TOY($m_X = 20$)	20000	20
TOY($m_X = 40$)	20000	40
AIDS	1156	11
JOBS	3212	17
TWINS	11400	39

hyper-parameters that correspond to the smallest MSE $(\widehat{Y}(1) - Y(1))^2$ on the training data, which indicates the optimal hyper-parameters for ϵ_{PEHE} on TOY($m_X = 20$). The optimal hyper-parameters for each dataset can be found in Table 5 in Appendix A.3.2.

A.3.3 Discussion on the Scalability to Arbitrary Forms of Treatments

It should be noted that our work can be naturally extended to arbitrary forms of treatments and has rigorous theoretical guarantees regarding the identifiability of true HTE in all strata, i.e., $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$. This way, by defining delayed response time $D(w)$ for all $w \in \mathcal{W}$ similarly and following a similar argument of our identifiability proof, and substitute $Y(0)$ and $Y(1)$ to $Y(w)$ for all $w \in \mathcal{W}$, the true HTE $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$ can be identified similarly. Moreover, in the proposed time-to-event based HTE problem setup with delayed responses, the outcome of interest has to be binary to ensure well-definiteness. Specifically, an event may either occur or not occur under any form of intervention (see the discussion in the previous paragraph), i.e., $Y(w) = 1$ or not $Y(w) = 0$. It is worth noting that only the former, i.e., $Y(w) = 1$, may be subject to delayed response, leading to the "false negative" samples. For the latter, $Y(w) = 0$, it is difficult to define a delayed response because this event never occurs (hence we let $D(w) = \infty$ for $Y(w) = 0$), and we will never observe "false positive" samples. To the best of our knowledge, this is the first work in the field of causal inference to consider the potential delayed response time $D(w)$ from intervention to outcome, and we theoretically prove the identifiability of true HTE in all strata. Considering the time it takes for an intervention to have an effect on an outcome, we believe this provides reasonable motivation in the causal inference community.

A.4 Extension to Non-binary Scenario

Our work can be naturally extended to non-binary treatments with the identifiability results of true HTE in all strata, i.e., $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$. By defining delayed response time $D(w)$ for all $w \in \mathcal{W}$ similarly and following a similar argument of our identifiability proof, and substitute $Y(0)$ and $Y(1)$ to $Y(w)$ for all $w \in \mathcal{W}$, the true HTE $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$ can be identified similarly. Moreover, in the proposed time-to-event based HTE problem setup with delayed responses, the outcome of interest has to be binary to ensure well-definiteness. Specifically, an event may either occur or not occur under any form of intervention (see the discussion in the previous paragraph), i.e., $Y(w) = 1$ or not $Y(w) = 0$. Only the former, i.e., $Y(w) = 1$, may be subject to delayed response, leading to the "false negative" samples. For the latter, $Y(w) = 0$, it is difficult to define a delayed

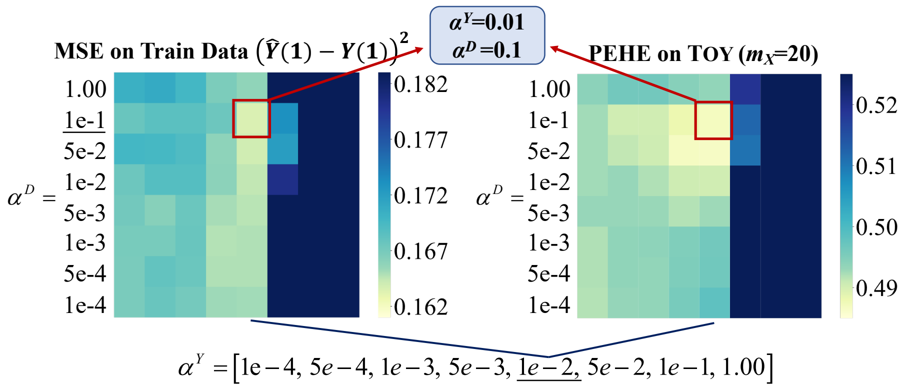


Figure 5: Hyper-Parameter Optimization: The smallest MSE on Train Data implies the best Hyper-Parameters. The optimal hyper-parameters are $\alpha^Y = 0.01, \alpha^{D=0.1}$ for TOY($m_X = 20$).

response because this event never occurs (hence we let $D(w) = \infty$ for $Y(w) = 0$), and we will never observe "false positive" samples. To the best of our knowledge, this is the first work in the field of causal inference to consider the potential delayed response time $D(w)$ from intervention to outcome, and we theoretically prove the identifiability of true HTE in all strata.

A.5 Datasets and Experiments

A.5.1 Datasets Used for Evaluation

Synthetic Datasets. Following the data generation process in Section 4.2, we generated data as follows. The observed covariates are generated from $X \sim \mathcal{N}(0, I_{m_X})$, where I_{m_X} denotes m_X-degree identity matrix. The observed treatment $W \sim \text{Bern}(\pi(X))$, where $\pi(X) = \mathbb{P}(W = 1 \mid X) = \sigma(\theta_W \cdot X)$, $\theta_W \sim U(-1, 1)$, and $\sigma(\cdot)$ denotes the sigmoid function. For the eventual potential outcomes, we generate the control outcome $Y(0) \sim \text{Bern}(\sigma(\theta_{Y0} \cdot X^2 + 1))$, and the treated outcome $Y(1) \sim \text{Bern}(\sigma(\theta_{Y1} \cdot X^2 + 2))$, where $\theta_{Y0}, \theta_{Y1} \sim U(-1, 1)$. In addition, we generate the potential response time $D(0) \sim \text{Exp}(\exp(\theta_{D0} \cdot X)^{-1})$, and $D(1) \sim \text{Exp}(\exp(\theta_{D1} \cdot X - b_D)^{-1})$, where $\theta_{D0}, \theta_{D1} \sim U(-0.1, 0.1)$, where $b_D = 0.5$ *controls the heterogeneity of response time functions*. The observation time is generated via $T \sim \text{Exp}(\lambda)$, where λ refers to the rate parameter of the exponential distribution. We set the rate parameter as $\lambda = 1$, i.e., the average observation time is $\bar{T} = \lambda^{-1} = 1$. Finally, the observed outcome is given as $\tilde{Y}^T(W) = W \cdot Y(1) \cdot \mathbb{I}(T \geq D(1)) + (1 - W) \cdot Y(0) \cdot \mathbb{I}(T \geq D(0))$, where $\mathbb{I}(\cdot)$ is the indicator function. From the data generation process described above, we sample $N = 20,000$ samples for training and $3,000$ samples for testing. We repeat each experiment 10 times to report the mean and standard deviation of the errors.

Real-World Datasets. In this paper, we use three wide-applied three widely-adopted real-world datasets: AIDS (`https://scikit-survival.readthedocs.io/` (Hammer et al., 1997; Norcliffe et al., 2023)), JOBS(`http://www.fredjo.com/` (LaLonde, 1986; Shalit et al., 2017)), and TWINS(`http://www.nber.org/data/` (Almond et al., 2005; Wu et al., 2022)). In Table 6 we provide details about the datasets used in our evaluation. The AIDS data collected between January 1996 and January 1997 involved 1,156 patients in 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the United States and Puerto Rico, and was used to study the impact and effectiveness of antiretroviral therapy on HIV-positive patients. The JOBS benchmark is widely used in the field of causal inference. It is built upon randomized controlled trials and aims to assess the effects of job training programs on employment status. The TWINS is derived from all twins born in the USA between the years 1989 and 1991, and is utilized to assess the influence of birth weight on mortality within one year.

Covariates X are obtained from AIDS, JOBS, and TWINS. Following the same procedure for generating synthetic datasets, we generate treatment W, potential outcomes $Y(0)$ and $Y(1)$, potential response times $D(0)$ and $D(1)$, observation time T and factual outcomes $\tilde{Y}^T(W)$. Then we randomly split the samples into training/testing with an 80/20 ratio with 10 repetitions.

Table 7: Performance comparison (MSE $\pm$ SD) on synthetic datasets with varying m_X.

	TOY ($m_X = 5$)		TOY ($m_X = 10$)	
Method	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}
T-learner	0.442 ± 0.028	0.028 ± 0.014	0.514 ± 0.036	0.028 ± 0.017
CFR	0.441 ± 0.029	0.029 ± 0.015	0.517 ± 0.037	0.025 ± 0.016
SITE	0.568 ± 0.039	0.029 ± 0.025	0.646 ± 0.077	0.026 ± 0.020
Dragonnet	0.457 ± 0.031	0.053 ± 0.037	0.499 ± 0.023	0.028 ± 0.024
CFR-ISW	0.463 ± 0.053	0.030 ± 0.022	0.602 ± 0.084	0.034 ± 0.024
DR-CFR	0.445 ± 0.033	0.040 ± 0.018	0.521 ± 0.044	0.032 ± 0.026
DER-CFR	0.462 ± 0.029	0.037 ± 0.020	0.540 ± 0.037	0.066 ± 0.043
CEVAE	0.590 ± 0.038	0.126 ± 0.028	0.661 ± 0.077	0.122 ± 0.039
GANITE	0.591 ± 0.036	0.149 ± 0.026	0.662 ± 0.075	0.147 ± 0.036
T-DF	<u>0.353 ± 0.057</u>	<u>0.022 ± 0.023</u>	<u>0.432 ± 0.013</u>	<u>0.017 ± 0.014</u>
CFR-DF	**0.329 ± 0.022**	**0.015 ± 0.013**	**0.404 ± 0.014**	**0.013 ± 0.009**
	TOY ($m_X = 20$)		TOY ($m_X = 40$)	
T-learner	0.593 ± 0.015	0.035 ± 0.014	0.677 ± 0.014	0.041 ± 0.010
CFR	0.588 ± 0.015	0.036 ± 0.017	0.678 ± 0.014	0.043 ± 0.011
SITE	0.716 ± 0.030	0.030 ± 0.017	0.760 ± 0.017	0.041 ± 0.014
Dragone	0.596 ± 0.016	0.034 ± 0.009	0.739 ± 0.021	0.041 ± 0.021
CFR-ISW	0.687 ± 0.033	0.056 ± 0.024	0.763 ± 0.030	0.070 ± 0.031
DR-CFR	0.633 ± 0.032	0.047 ± 0.035	0.754 ± 0.028	0.043 ± 0.022
DER-CFR	0.665 ± 0.030	0.086 ± 0.032	0.754 ± 0.025	0.053 ± 0.043
CEVAE	0.722 ± 0.030	0.098 ± 0.016	0.762 ± 0.028	0.078 ± 0.014
GANITE	0.717 ± 0.029	0.081 ± 0.016	0.762 ± 0.027	0.066 ± 0.015
T-DF	<u>0.529 ± 0.011</u>	<u>0.018 ± 0.013</u>	<u>0.633 ± 0.008</u>	<u>0.018 ± 0.011</u>
CFR-DF	**0.498 ± 0.021**	**0.017 ± 0.010**	**0.612 ± 0.007**	**0.012 ± 0.007**

A.5.2 MORE EXPERIMENTS ON VARYING FEATURE DIMENSIONS

To evaluate our CFR-DF on a wide range of scenarios, given $b_D = 0.5$, we further tune the number of features by varying the dimension $m_X \in \{5, 10, 20, 40\}$, named the dataset as TOY ($m_X = 5$), TOY ($m_X = 10$), TOY ($m_X = 20$), and TOY ($m_X = 40$), respectively.

Performance Comparison. Table 7 presents a comprehensive performance comparison between our proposed method and the baselines in estimating the Heterogeneous Treatment Effect (HTE) on the eventual outcome, considering varying feature dimensions. The optimal and second-optimal performances are indicated as **bold** and <u>underlined</u>, respectively. Consistent with the observations from Table 2, our CFR-DF consistently outperforms the baselines, demonstrating its efficacy in addressing the label noise arising from delayed responses. In contrast, previous methods that do not consider delayed responses often yield biased estimates of HTE. Additionally, the T-DF method without using balancing regularization slightly degrades the performance compared to CFR-DF, due to the inability to resolve the confounding bias from covariate shift. Overall, our method achieves significant reductions in the ϵ_{PEHE} and ϵ_{ATE}. Specifically, comparing CFR-DF to the optimal traditional causal method in the ϵ_{PEHE} and ϵ_{ATE}, we observe reductions of 25% and 46% in TOY ($m_X = 5$), 21% and 48% in TOY ($m_X = 10$), 15% and 43% in TOY ($m_X = 20$), and 10% and 70% in TOY ($m_X = 40$), respectively. These results highlight the superior performance of CFR-DF compared to the baselines and its scalability to different feature dimensions, further emphasizing its potential for accurate and robust estimation of HTE in various practical settings.