Individual Treatment Prescription Effect Estimation in a Low Compliance Setting

Thibaud Rahier, Amélie Héliou, Matthieu Martin, Christophe Renaudin, Eustache Diemert

Individual Treatment Prescription Effect Estimation in a Low Compliance Setting

Thibaud Rahier
CAIL

Amélie Héliou
CAIL

Matthieu Martin
CAIL

Christophe Renaudin
CAIL

Eustache Diemert
CAIL

September 14, 2021

Abstract

Individual Treatment Effect (ITE) estimation is an extensively researched problem, with applications in various domains. We model the case where there exists heterogeneous non-compliance to a randomly assigned treatment, a typical situation in health (because of non-compliance to prescription) or digital advertising (because of competition and ad blockers for instance). The lower the compliance, the more the effect of treatment prescription − or individual prescription effect (IPE) − signal fades away and becomes harder to estimate. We propose a new approach for the estimation of the IPE that takes advantage of observed compliance information to prevent signal fading. Using the Structural Causal Model framework and *do-calculus*, we define a general mediated causal effect setting and propose a corresponding estimator which consistently recovers the IPE with asymptotic variance guarantees. Finally, we conduct experiments on both synthetic and real-world datasets that highlight the benefit of the approach, which consistently improves state-of-the-art in low compliance settings.

1 Introduction

Individual Treatment Effect (ITE) estimation is an important task in various applications such as healthcare (Foster et al., 2011), online advertising (Diemert et al., 2018) or socio-economics (Xie et al., 2012). As it is often the case in practice, we assume that we cannot directly control the treatment intake but only the treatment prescription: we therefore focus on the Individual Prescription Effect (IPE), which designates the effect of the *treatment prescription* P on the *outcome* Y for an individual described by *covariates* X (*c.f.* Equation (1), assuming random prescription assignment).

$$IPE(x) = \mathbb{E}[Y|X = x, P = 1] - \mathbb{E}[Y|X = x, P = 0] \tag{1}$$

We also assume that we observe the *treatment intake* (or equivalently, the compliance to prescription), denoted by T, which acts as a *mediator* of the causal effect of P on Y. The effect of T on Y will henceforth be referred to as ITE.

Table 1 illustrates various settings in which non-compliance to treatment prescription may occur (Gordon et al., 2019; Jin et al., 2008). This happens for instance when individuals have the choice not to abide by the prescription or if there exists conflicting interests. Of course one can choose to focus the study on actually treated individuals only (ITE). But from a decision making point of view it often makes sense to consider that future treatment decisions need to take into account the possibility of non-compliance so as to accurately predict future expected outcomes. For example a policy maker would want to take into account that not all individuals would abide by the new policy (as can be estimated from a pilot study) to predict the expected impact of a roll-out of said policy.

Table 1: Examples of Covariates (X), Outcome (Y), Treatment Prescription (P) and Evidence of Treatment Acceptance (T)

Var.	Medicine	Online adv.	Job training
X	patient info	purchase history	schooling
P	drug prescription	bid placement	training offer
T	drug intake	ad displayed	training done
Y	recovery	sale/visit	employment

Now we argue that IPE estimation can be hampered by non-compliance. Indeed, individuals who did not actually receive the treatment contribute only through noise to the default IPE estimator, as their observed outcome is not effectively influenced. The variance of the IPE estimator therefore increases as the compliance level decreases.

Besides, individual treatment (whether prescription or intake) effect models are often considered as prescriptive tools. Indeed, treatment effect predictions can be used in order to target treatment to individuals for which it is the most beneficial (Devriendt et al., 2018; Radcliffe and Surry, 2011). This calls for an evaluation metric that measures by how much the average treatment effect would have been improved, had the treatment been targeted not by a random instrument, but by the predictions of the considered model instead. For that purpose Rzepakowski and Jaroszewicz (2012), Rzepakowski and Jaroszewicz (2010), and Radcliffe and Surry (2011) have proposed the *Area Under the Uplift Curve* (AUUC) metric that sums the benefits over individuals ranked by predictions. An interesting property of this metric is that it can be used on real data for which we observe a given individual in either treated or untreated conditions but never both.

Confronted with the challenges of (i) learning IPE models in conditions of (possibly high) non-compliance and (ii) evaluating them as prescriptive tools, we pose the problem in the setting of causal inference and derive an IPE estimator that takes advantage of observed compliance. Our main contributions are as follows.

1. Formalization of IPE estimation in a setting of observed compliance described using structural causal models, which extends to more general cases of mediated causal effect estimation (Section 3)

2. Proposition of a meta-estimator (in which can be plugged any ITE estimator) for IPE estimation in presence of non-compliance, proof of consistency (Section 4.1) and asymptotic variance properties (Section 4.2)

3. Thorough empirical evaluation of this meta-estimator on synthetic and real world datasets (Section 5)

2 RELATED WORK

We review three main domains that are concerned with research questions similar to our work: ITE modeling, non compliance in causal inference and evaluation metrics for ITE modeling.

First, ITE models are a pervasive concept in different research fields such as marketing - under the name uplift models (Radcliffe and Surry, 2011), statistics - as conditional average treatment effect estimators (Künzel et al., 2019) or econometrics - heterogeneous treatment effect models (Jacob et al., 2019; Wager and Athey, 2018). A simple yet highly scalable approach consists in learning a regression of Y on X separately in both treatment ($P = 1$) and control ($P = 0$) populations and return the difference, known as T-learner studied in Künzel et al. (2019) or "Two Models" studied in Radcliffe and Surry (2011). A variation of this approach with larger model capacity have been proposed through a shared representation (SDR) for the treatment and control group in Betlei et al. (2018). Also, a prolific series of increasingly performing algorithm targeted towards the recovery of the ITE in observational settings have been proposed (Yoon et al., 2018; Zhang et al., 2020) a sub-category of which tackles the use of decision trees and random forests in a causal

inference framework (Athey and Imbens, 2016; Athey et al., 2019). Further in the same vein and building on work done on double machine learning (Chernozhukov et al., 2018), Oprescu et al. (2019) generalize the idea of causal forests, allowing for high-dimensional confounding. Other work such as Kuang et al. (2017) handle confounding bias by confounder balancing algorithms. Finally, another recent trend is to study theoretical limits in ITE estimation and especially generalization bounds (Alaa and Schaar, 2018).

Then regarding the concept of non compliance, many algorithms have been studied in order to recover (individual) causal effects in non-compliance settings, however to our knowledge, works tackling this problem (*e.g.* Gordon et al. (2019)) focus solely on effect of the treatment *intake* T − and not the treatment *prescription* P − on the outcome Y. In that context, the effect of P on Y is sometimes studied in an *instrumental variable* framework to recover the effect of T on Y (Imbens and Angrist, 1994). Our work however focuses on estimating the effect of the treatment prescription P on the variable Y, taking advantage of the observed treatment intake T, in the spirit of individual *intention-to-treat* (ITT) analysis (Gupta, 2011) or per-protocol effect (Hernán et al., 2017). The idea of taking advantage of a mediation variable for causal effect estimation has triggered many recent works (Yin and Hong, 2019), in settings sometimes similar to ours (Loeys et al., 2015). However, the associated assumptions (T and Y are unconfounded) are more restrictive than the ones we propose, as we allow all the covariates (X) to be confounders between the mediation variable T (representing the treatment intake) and Y, and state a result holding for individual (and not only average) treatment effect.

Finally, treatment effect estimators performance is typically done using synthetic data, in which a pointwise error measure − Precision Estimation of Heterogeneous Effect (PEHE) (Hill, 2011) − is available. However in real world cases, the *fundamental problem of causal inference* states that the ground truth of individual treatment effect cannot be observed (since an individual is either treated or untreated but never both at the same time), preventing to use such metrics beyond simulation studies. Since our main motivation is to determine which individuals are good candidates for treatment *prescription*, we choose to evaluate the performance of our estimators on real data using the AUUC metric, which evaluates the ranking of individuals implied by corresponding ITE or IPE predictions. One can view the resulting measure as a prediction of the expected benefit of prescribing treatment according to the model prediction instead of a random uniform prescription. Overall, AUUC has been used in recent years in machine learning research to evaluate baseline ITE models vs SDR (Diemert et al., 2018), flavors of Support Vector Machines for ITE estimation (Kuusisto et al., 2014) or direct treatment policy optimization (Yamane et al., 2018).

3 FRAMEWORK

We briefly recall causality notions used throughout the paper, such as *structural causal model*, *causal graph*, *intervention* and *valid adjustment set* first introduced in Pearl (2009), and more recently presented in Peters et al. (2017).

Definition 1 *Structural Causal Model (SCM) A SCM of variables* $\mathbf{X} = \{X_1, \dots, X_d\}$ *is an object* $\mathfrak{C} := (\mathbb{S}, \mathbb{P}_{\mathbf{N}})$ *where:*
(1) $\mathbb{S}$ *is a set of d structural assignments* $X_i = f_i(\mathbf{PA}_i, N_i)$*, with the* f_i*'s deterministic functions and* $\mathbf{PA}_i$ *the set of parents (direct causes) of* X_i*.*
(2) $\mathbb{P}_{\mathbf{N}} = \mathbb{P}_{N_1,\dots,N_d}$ *is a joint distribution over the noise variables* $\{N_i\}_{1\leq d}$*, which we require to be jointly independent.*

As shown in Chapter 6 of Peters et al. (2017), a SCM $\mathfrak{C}$ induces a unique *causal graph* $\mathcal{G}_{\mathfrak{C}}$ − defined as the directed acyclic graph (DAG) obtained by creating a vertex for each X_i and drawing directed edges from each $\mathbf{PA}_i$ to X_i (thus justifying the term 'parents' for the sets $\mathbf{PA}_i$) − and a unique *entailed distribution* $\mathbb{P}^{\mathfrak{C}}$ over variables $X_1, \dots, X_d$ such that for each i, $X_i = f_i(\mathbf{PA}_i, N_i)$ in distribution.

A SCM can be used to define *interventional distributions*. A (hard) intervention $do(X_i = x_i)$ is a forced assignment of variable X_i to the value x_i, which implies a change in the distribution of the variables $X_1, \dots, X_d$. Formally, an intervention $do(X_i = x_i)$ is equivalent to modifying $\mathfrak{C}$ in two ways: (1) change

the structural assignment $X_i = f_i(X_i, \mathbf{PA}_i)$ to $X_i = x_i$ in $\mathbb{S}$ and (2) replace X_i by x_i in all other structural assignments implying X_i in $\mathbb{S}$.
We will denote by $\mathbb{P}^{\mathfrak{C};do(X_i:=x_i)}$ the probability distribution entailed by the SCM $\mathfrak{C}$ under the intervention $do(X_i := x_i)$.

Let $X_i, X_j \in \mathbf{X}$, $\mathbf{Z} \subseteq \mathbf{X} \backslash \{X_i, X_j\}$ is called a *valid adjustment set* (VAS) for the ordered pair (X_i, X_j), if

$$\mathbb{P}^{\mathfrak{C};do(X_i)}(X_j) = \sum_{\mathbf{z}} \mathbb{P}^{\mathfrak{C}}(X_j | X_i, \mathbf{z}) \mathbb{P}^{\mathfrak{C}}(\mathbf{z}),$$

where the sum is over the range of $\mathbf{Z}$.

Notations For sake of compactness, we use the following notations for any binary variable W and multi-dimensional variable X: $\mathbb{P}(W) \triangleq \mathbb{P}(W = 1)$, $\mathbb{P}(\overline{W}) \triangleq \mathbb{P}(W = 0)$, $\mathbb{P}(x) \triangleq \mathbb{P}(X = x)$, $\mathrm{do}(W) \triangleq \mathrm{do}(W := 1)$.

3.1 Non-compliance setting

The setting of non-compliance we consider in this work is entirely defined by a SCM of variables X, P, T, Y, U for which example values where proposed in Table 1: X, belonging to a multi-dimensional space $\mathcal{X}$, contains the individual's descriptive features, or *covariates* (by simplicity, we will confuse individuals and their covariates, referring for example to 'an individual x'), P is the binary treatment *prescription* variable, Y is the binary *outcome* variable, T is the binary treatment *intake* (or acceptation) − which acts as a *mediator* of the causal effect of P on Y − and U represents (allowed) *unobserved confounders* between X and Y.

In what follows, we define the structural causal model $\mathfrak{C} = (\mathbb{S}, \mathbb{P}_{\mathbf{N}})$, which is henceforth assumed to represent the causal mechanisms underlying the variables of interest in this work. The associated causal graph $\mathcal{G}_{\mathfrak{C}}$ is given in Figure 1.

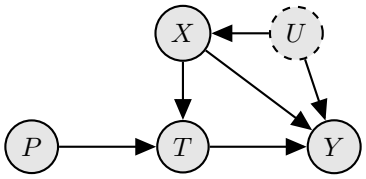

Figure 1: **Causal Graph $\mathcal{G}_{\mathfrak{C}}$ Induced by SCM $\mathfrak{C}$**

$\mathbb{S}$ is defined in Equations (2):

$$\begin{aligned} P &= \tilde{N}_P \\ U &= N_U \\ X &= f_X(U, N_X) \\ T &= f_T(X, N_T) \times P \\ Y &= f_Y(X, T, U, N_Y). \end{aligned} \tag{2}$$

$\mathbb{P}_{\mathbf{N}}$ satisfies the following mild conditions: N_U, N_X, N_T, N_Y are noise consistent with variables definitions, and $\tilde{N}_P$ is distributed according to a Bernoulli distribution with parameter $p = \mathbb{P}^{\mathfrak{C}}(P)$, consistent with a *randomized controlled experiment* setting.

In the next proposition, we list four assumptions implied by $\mathfrak{C}$ about the variables of interest.

Proposition 1 *The SCM* $\mathfrak{C}$ *defined in Equations* (2) *implies the following assumptions on variables* P*,* T*,* Y *and* X*:*

$$\begin{aligned} &\textit{(Randomized prescription)} && P \perp\!\!\!\perp X \\ &\textit{(Exclusive treatment effect)} && P \perp\!\!\!\perp Y \mid \{X, T\} \\ &\textit{(One-sided non-compliance)} && P = 0 \Rightarrow T = 0 \\ &\textit{(Valid covariate adjustment)} && \{X\} \textit{ is a VAS for } (T, Y) \end{aligned} \tag{3}$$

The proof of Proposition 1 relies on the notions of valid adjustment set, its relation to the back-door criterion, the notion of d-separation and the Markov property (Definitions 6.1 and 6.38, Propositions 6.21 and 6.41 of Peters et al. (2017)). More details are given in the supplement.

Detailed review of assumptions mentioned in Proposition 1 . We now review in details the assumptions presented in Equation (3) and their connections to applications and usual theoretical settings.
- *Randomized prescription* translates the fact that each individual is randomly allocated to a prescription ($P = 1$) or a non-prescription / control ($P = 0$) group, allowing us to infer the causal effect of P on Y without worrying about confounding bias. This is a typical situation in experiments such as A/B tests or randomized controlled trials, which motivate increasingly substantial research (Xie et al., 2018).
- *Exclusive treatment effect* captures the fact that the prescription to treatment P has an effect on the outcome Y only if the treatment is actually taken $T = 1$. More generally, this assumption translates the fact that the effect of P on Y is fully mediated by T. This is typically assumed to be the case in online advertising, in which case P is the allocation to a given advertising engine, and T is the actual exposure to advertisement from this engine, as in the study of Gordon et al. (2019). This assumption is typically satisfied in the instrumental variable setting where P plays the role of the instrument for the estimation of the effect of T on Y (Syrgkanis et al., 2019). There are essential differences with our work, notably (i) we are interested in the effect of P on Y (IPE) and not in the effect of T on Y, and (ii) *one-sided non-compliance* is not assumed in typical IV settings.
- *One-sided non-compliance* defined in Gordon et al. (2019) states that $P = 0 \implies T = 0$, but that $P = 1 \not\Longrightarrow T = 1$. In other words, treatment is only taken ($T = 1$) in cases where it was originally assigned ($P = 1$).
- *Valid covariate adjustment*: every confounder for the effect of T on Y is observed (and contained in X). This assumption is classically made in causal inference works (Alaa and Schaar, 2018; Alaa and Van Der Schaar, 2019) and is sometimes referred to as *ignorability* (Shalit et al., 2017). Under the potential outcome framework (Rubin, 2005), this assumption is equivalent to

$$Y(T = 1), Y(T = 0) \perp\!\!\!\perp T|X.$$

Strong ignorability (where we additionally assume that $0 < \mathbb{P}(T|x) < 1$ for all x) has been proven to be sufficient to recover the individual causal effect of T on Y (Imbens and Wooldridge, 2009; Pearl, 2017).

3.2 IPE in presence of non-compliance

Notations For all $x \in \mathcal{X}$, we formally define the individual prescription effect $\tau^{IPE}(x)$, the individual treatment effect $\tau^{ITE}(x)$ and the *individual compliance* probability $\gamma(x)$ as follows:

$$\begin{aligned} \tau^{IPE}(x) &= \mathbb{P}^{\mathfrak{C};do(P)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{P})}(Y|x), \\ \tau^{ITE}(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x), \\ \gamma(x) &= \mathbb{P}^{\mathfrak{C};do(P)}(T|x). \end{aligned} \tag{4}$$

We also define the relative ITE $\beta(x)$ and relative IPE $\alpha(x)$ as:

$$\begin{aligned} \alpha(x) &= \left(\mathbb{P}^{\mathfrak{C}}(Y|P, x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{P}, x)\right) / \mathbb{P}^{\mathfrak{C}}(Y|\overline{P}, x), \\ \beta(x) &= \left(\mathbb{P}^{\mathfrak{C}}(Y|T, x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{T}, x)\right) / \mathbb{P}^{\mathfrak{C}}(Y|\overline{T}, x). \end{aligned}$$

We propose to exploit the mediation variable T by splitting the P to Y path into two *subpaths* (from P to T and from T to Y), both with a higher signal-to-noise ratio. Indeed, under $\mathfrak{C}$ we can insert T into the $\mathbb{P}^{\mathfrak{C};do(P)}(Y|x)$ expression as presented in the next lemma.

Lemma 1 *Assuming $\mathfrak{C}$, and for any $x \in \mathcal{X}$, the positive outcome probability under treatment, $\mathbb{P}^{\mathfrak{C};do(P)}(Y|x)$, can be written as follows:*

$$\mathbb{P}^{\mathfrak{C};do(P)}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T}) \ + \ \tau^{ITE}(x)\gamma(x). \tag{5}$$

Proof.
Assuming the SCM $\mathfrak{C}$ truly describes the relationships between P, X, T, Y, we have:

$$\begin{aligned}
&\mathbb{P}^{\mathfrak{C};do(P)}(Y|x) = \mathbb{P}^{\mathfrak{C};do(P)}(Y,T|x) + \mathbb{P}^{\mathfrak{C};do(P)}(Y,\overline{T}|x) \\
&= \mathbb{P}^{\mathfrak{C};do(P)}(Y|x,T)\mathbb{P}^{\mathfrak{C};do(P)}(T|x) + \mathbb{P}^{\mathfrak{C};do(P)}(Y|x,\overline{T})\mathbb{P}^{\mathfrak{C};do(P)}(\overline{T}|x) \\
&= \mathbb{P}^{\mathfrak{C}}(Y|x,T)\mathbb{P}^{\mathfrak{C};do(P)}(T|x) + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T})\underbrace{\mathbb{P}^{\mathfrak{C};do(P)}(\overline{T}|x)}_{1-\mathbb{P}^{\mathfrak{C};do(P)}(T|x)} \\
&= \mathbb{P}^{\mathfrak{C}}(T|x,P)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,T) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T})\Big) + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T}),
\end{aligned}$$

where we used assumptions described Equation 3, namely:

- $\mathbb{P}^{\mathfrak{C};do(T)}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|T,x)$ (*Valid covariate adjustment*),
- $\mathbb{P}^{\mathfrak{C};do(P)}(\cdot|x,\cdot) = \mathbb{P}^{\mathfrak{C}}(\cdot|x,P,\cdot)$ (*Randomized prescription*),
- $\mathbb{P}^{\mathfrak{C}}(Y|x,P,T) = \mathbb{P}^{\mathfrak{C}}(Y|x,T)$ (*Exclusive treatment effect*),

and the claim follows. ■

In the illustrative setting of computational advertising, where $\mathfrak{C}$ is typically satisfied, Lemma 1 states that the *conversion rate of a web user given they were targeted by an advertising campaign* is equal to *their organic conversion rate*, plus the product of the *effect of ad exposure on this user* times the *probability that this user was effectively exposed*.

In Proposition 2, we present a result linking the IPE, the ITE and the individual compliance:

Proposition 2 *Assuming $\mathfrak{C}$ and for any $x \in \mathcal{X}$, the IPE decomposes as follows:*

$$\tau^{IPE}(x) = \tau^{ITE}(x)\gamma(x) \tag{6}$$

Proof
We have an analogous version of (5) for the term $\mathbb{P}^{\mathfrak{C};do(\overline{P})}(Y|x)$:

$$\mathbb{P}^{\mathfrak{C};do(\overline{P})}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T}) + \mathbb{P}^{\mathfrak{C}}(T|x,\overline{P})\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,T) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T})\Big).$$

Since $\overline{P} \Rightarrow \overline{T}$ (*One-sided non-compliance* assumption), we get that, $\forall x \in \mathcal{X}$, $\mathbb{P}^{\mathfrak{C}}(T|x,\overline{P}) = 0$, and finally:

$$\mathbb{P}^{\mathfrak{C};do(\overline{P})}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T}).$$

Then:

$$\begin{aligned}
\tau^{IPE}(x) &= \mathbb{P}^{\mathfrak{C};do(P)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{P})}(Y|x) \\
&= \mathbb{P}^{\mathfrak{C}}(T|x,P)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,T) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T})\Big).
\end{aligned}$$

which completes the proof.

■

In intuitive terms, Proposition 2 states that the effect of treatment prescription on a given individual is equal to the effect of treatment intake on this individual, multiplied by their compliance. In the online advertising setting, this means that the *effect of targeting a user with an advertising campaign* is equal to the *effect of effectively exposing them to advertisement* times *the probability of the campaign succeeding in doing so*.

4 PROPOSED APPROACH

4.1 Definition and basic properties

The expression proven in Proposition 2 calls for a novel way to estimate $\tau^{IPE}(x)$, by first estimating separately both factors $\tau^{ITE}(x)$ and $\gamma(x)$, then multiplying these estimators to form a *compliance-aware individual prescription effect* (C-IPE) estimator.
Formally, let $\hat{\tau}^{ITE}$ be an estimator of τ^{ITE} and $\hat{\gamma}$ be an estimator of γ. We define the associated 'plug-in' C-IPE estimator $\hat{\tau}^{C-IPE}$, for any x, as:

$$\hat{\tau}^{C-IPE}(x) = \hat{\tau}^{ITE}(x)\hat{\gamma}(x). \tag{7}$$

In practice, $\hat{\tau}^{ITE}$ may be any individual treatment effect estimator. Indeed, under $\mathfrak{C}$, the individual causal effect of T on Y given X is identifiable since $\{X\}$ is a valid adjustment set for (T, Y) as explained in Section 3. C-IPE is not an estimator per say, but rather a *meta-estimator*: a black-box in which any classical ITE and compliance estimators can be plugged-in.

Assuming that for all x, $\hat{\tau}^{ITE}(x)$ and $\hat{\gamma}(x)$ are *consistent* estimators of resp. $\tau^{ITE}(x)$ and $\gamma(x)$, Proposition 2 then ensures that $\hat{\tau}^{C-IPE}(x)$ is a *consistent* estimator of the associated $\tau^{IPE}(x)$.

4.2 Asymptotic variance properties

Thanks to its expression as a function of an ITE estimator, the C-IPE estimator *focuses* on the individuals who actually received treatment, who are exclusively contributing to the signal (*Exclusive treatment effect* assumption in Proposition 1). We therefore expect the C-IPE estimator to have lower variance than any standard IPE estimator which does not exploit observable compliance.
Comparing C-IPE and IPE estimators is all the more fair than we use an analogous version of the standard IPE estimator $\hat{\tau}^{IPE}$ for the ITE estimator $\hat{\tau}^{ITE}$ which is plugged in the C-IPE estimator $\hat{\tau}^{C-IPE}(x)$. Doing so is feasible since the *valid covariate adjustment* assumption guarantees that both τ^{IPE} and τ^{ITE} are identifiable and therefore estimable with standard treatment effect estimators. We refer to this approach as *symmetrically formed estimators comparison*, and use it to conduct our experiments in Section 5.
In the following proposition, we compare the asymptotic variance of estimators $\hat{\tau}^{C-IPE}$ and $\hat{\tau}^{IPE}$ in the following simple yet realistic setting:

Single-stratum setting We focus on the IPE estimation for a single value x_0 of X, for which we assume to observe n *i.i.d.* samples $\{(x_0, P_i, T_i, Y_i)\}_{1\leq i\leq n}$. All stated results generalise to any stratum $S \subset \mathcal{X}$ for which there exists $x_S \in \mathcal{X}$ such that the reducted feature variable $X' \triangleq x_S \mathbb{I}_{\{X\in S\}} + X\mathbb{I}_{\{X\notin S\}}$ still defines a valid adjustment set for (T, Y).

Notations Consistently with notations presented in Equations (4), $\alpha(x_0)$, $\beta(x_0)$ refer respectively to the relative IPE and relative ITE in stratum $\{X = x_0\}$ (and are assumed to be positive in this illustrative setting), and we denote $\hat{\gamma}(x_0)$, $\hat{\tau}^{IPE}(x_0)$ and $\hat{\tau}^{ITE}(x_0)$ respectively the maximum-likelihood estimator (MLE) of $\gamma(x_0)$, and the MLE-based two-model estimators (difference of two MLE estimators) of $\tau^{IPE}(x_0)$ and $\tau^{ITE}(x_0)$. We define the associated C-IPE estimator as $\hat{\tau}^{C-IPE}(x_0) \triangleq \hat{\gamma}(x_0)\hat{\tau}^{ITE}(x_0)$. Note that we use the MLE-based Two-Model estimator for both $\hat{\tau}^{ITE}$ and $\hat{\tau}^{IPE}$, thus satisfying the *symmetrically formed estimators comparison* setting. Lastly, we denote $p_1(x_0) = \mathbb{P}^{\mathfrak{C}}(Y|P, x_0)$.
In the following Proposition, we present an asymptotic bound for the ratio of the standard deviation (sd) of C-IPE and IPE estimators.

Proposition 3 *Under $\mathfrak{C}$ defined in Section 3.1 with $\mathbb{P}^{\mathfrak{C}}(P) = \frac{1}{2}$, assuming we observe n i.i.d. samples in stratum $\{X = x_0\}$, we have:*

$$\lim_{n\to\infty} \frac{sd(\hat{\tau}^{C-IPE})}{sd(\hat{\tau}^{IPE})} \leq \sqrt{\left(\frac{2(1+\beta)}{(1-p_1)(2+\alpha)}\right)\gamma} \tag{8}$$

where we dropped the reference to x_0 for clarity.

This theoretical bound shows that the ratio of standard deviations of C-IPE and IPE estimators gets smaller when the compliance factor γ decreases, which was expected. Of course the $\mathbb{P}^{\mathfrak{C}}(P) = \frac{1}{2}$ assumption is not necessary to obtain such a bound but is simply there to keep the result as concise as possible.

Before proving Proposition 3, let us note that additional assumptions need to be made on β and p_1 in order to derive an informative bound in practice. In real-world datasets (presented in Section 5), we consistently observe $\hat{\beta}(x) \leq 12$ and $\hat{p}_1(x) \leq 0.05$, with estimators described in Section 5. Under such assumptions on the orders of magnitude of β and p_1, we can derive a useful-in-practice bound, as presented in the following remark.

Remark 1 *If we additionally assume $\beta \leq 12$, and $p_1 \leq 0.05$ we have:*

$$\lim_{n\to\infty} \frac{sd(\hat{\tau}^{C-IPE})}{sd(\hat{\tau}^{IPE})} \leq 4\sqrt{\gamma} \tag{9}$$

where we dropped the reference to x_0 for clarity.

This bound is derived from loose upper-bounds on β and p_1 and is only presented for illustrative purposes. It is however still informative in case of low compliance γ. For instance, if $\gamma \leq 10^{-2}$, Equation (9) reveals that the asymptotic standard deviation of $\hat{\tau}^{C-IPE}$ is more than twice smaller than the one of $\hat{\tau}^{IPE}$.

We now provide a detailed sketch of proof for the bound provided by (8). Proposition 3.

Proof of Proposition 3.
The proof is split in four steps: (1) Maximum-Likelihood and treatment effect estimators, (2) Variance of estimators derivation, (3) Variances upper and lower bounds derivation and (4) Wrap up. Every random quantity is henceforth implicitly considered to be 'with respect to x_0'.

1. Maximum-Likelihood and treatment effect estimators. We have n *i.i.d.* samples $\{(P_i, T_i, Y_i)\}_{1\leq i\leq n}$ of variables (P, T, Y), and that we suppose $\mathbb{P}^{\mathfrak{C}}(P) = \frac{1}{2}$, and define the following compact notations: $t = \mathbb{P}^{\mathfrak{C}}(P)$, $\gamma = \mathbb{P}^{\mathfrak{C}}(T|P)$, $p_0 = \mathbb{P}^{\mathfrak{C}}(Y|\overline{P})$, $p_1 = \mathbb{P}^{\mathfrak{C}}(Y|P)$, $q_0 = \mathbb{P}^{\mathfrak{C}}(Y|\overline{T})$ and $q_1 = \mathbb{P}^{\mathfrak{C}}(Y|T)$.

Corresponding MLEs $\hat{p_0}$, $\hat{p_1}$, $\hat{q_0}$, $\hat{q_1}$, $\hat{t}$ and $\hat{\gamma}$ are given empirical frequencies, for example:

$$\hat{t} = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad\qquad \hat{p}_1 = \frac{1}{\sum_{i=1}^{n} P_i}\sum_{i=1}^{n} P_i Y_i$$

Therefore, direct estimators (2-model) for τ^{IPE}, τ^{ITE} are given by

$$\hat{\tau}^{IPE} = \hat{p}_1 - \hat{p}_0 \qquad\qquad \hat{\tau}^{ITE} = \hat{q}_1 - \hat{q}_0$$

and the corresponding τ^{C-IPE} estimator therefore writes:

$$\hat{\tau}^{C-IPE} = (\hat{q}_1 - \hat{q}_0)\hat{\gamma}.$$

2. Variance of estimators derivation

2.a. $\hat{\tau}^{IPE}$ variance derivation

For any random variables X, Y, we remind that:

$$\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathrm{Var}(X|Y)]. \tag{10}$$

Which applied with $X = \hat{\tau}^{IPE} = \hat{p}_1 - \hat{p}_0$ and $Y = \{P_k\}_k := \{P_1, \ldots, P_n\}$, gives:

$$\mathrm{Var}(\hat{\tau}^{IPE}) = \underbrace{\mathbb{E}\left[\mathrm{Var}(\hat{\tau}^{IPE}|\{P_k\}_k)\right]}_{\text{(A)}} + \underbrace{\mathrm{Var}\left[\mathbb{E}(\hat{\tau}^{IPE}|\{P_k\}_k)\right]}_{\text{(B)}}. \tag{11}$$

To compute Ⓐ, we remark that the term $\mathrm{Var}(\hat{\tau}^{IPE}|\{P_k\}_k)$ decomposes as:

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^{IPE}|\{P_k\}_k) = \mathrm{Var}(\hat{p}_1|\{P_k\}_k) + \mathrm{Var}(\hat{p}_0|\{P_k\}_k) \\ -2\mathrm{Cov}(\hat{p}_1, \hat{p}_0|\{P_k\}_k).\end{aligned} \tag{12}$$

Straightforward derivations bring

$$Ⓐ = \mathbb{E}\left[\frac{1}{\sum_i P_i}\right] p_1(1-p_1) + \mathbb{E}\left[\frac{1}{\sum_i (1-P_i)}\right] p_0(1-p_0). \tag{13}$$

To compute Ⓑ, we begin by writing

$$\mathbb{E}(\hat{\tau}^{IPE}|\{P_k\}_k) = \underbrace{\mathbb{E}(\hat{p_1}|\{P_k\}_k)}_{p_1} - \underbrace{\mathbb{E}(\hat{p_0}|\{P_k\}_k)}_{p_0}.$$

Therefore $\mathbb{E}(\hat{\tau}^{IPE}|\{P_k\}_k) = p_1 - p_0$ is constant relatively to $\{T_k\}_k$, and we have Ⓑ $= 0$. Combining this with Equations (11) and (13) we end up with:

$$\mathrm{Var}(\hat{\tau}^{IPE}) = \mathbb{E}\left[\frac{1}{\sum_i P_i}\right] p_1(1-p_1) + \mathbb{E}\left[\frac{1}{\sum_i (1-P_i)}\right] p_0(1-p_0). \tag{14}$$

2.B. $\hat{\tau}^{C-IPE}$ VARIANCE DERIVATION

Using Equation (10) with $X = \hat{\tau}^{C-IPE} = \hat{\gamma}(\hat{q}_1 - \hat{q}_0)$ and $Y = \{P_k, T_k\}_k$ we may write:

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^{C-IPE}) = \mathbb{E}\left[\mathrm{Var}(\hat{\tau}^{C-IPE}|\{P_k, T_k\}_k)\right]\} = Ⓒ \\ + \mathrm{Var}\left[\mathbb{E}(\hat{\tau}^{C-IPE}|\{P_k, T_k\}_k)\right]\} = Ⓓ.\end{aligned} \tag{15}$$

To compute Ⓒ, we use $\hat{\tau}^{C-IPE} = \hat{\gamma}(\hat{q}_1 - \hat{q}_0)$, and remark that $\mathbb{E}[\hat{\gamma}|\{P_k, T_k\}_k] = \hat{\gamma}$, which brings:

$$\mathrm{Var}(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{P_k, T_k\}_k) = \hat{\gamma}^2 \mathrm{Var}((\hat{q}_1 - \hat{q}_0)|\{P_k, T_k\}_k)$$

By analogy with Equation (14), we now have

$$\mathrm{Var}(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{P_k, T_k\}_k) = \frac{\hat{\gamma}^2}{\sum_i T_i} q_1(1-q_1) + \frac{\hat{\gamma}^2}{\sum_i (1-T_i)} q_0(1-q_0). \tag{16}$$

Injecting $\hat{\gamma} = \frac{\sum_i T_i}{\sum_i P_i} = \frac{\sum_i T_i P_i}{\sum_i P_i}$ in (16) and taking the expectancy we finally have

$$Ⓒ = \mathbb{E}\left[\frac{\sum_i T_i}{\left(\sum_i P_i\right)^2}\right] q_1(1-q_1) + \mathbb{E}\left[\frac{\left(\sum_i T_i\right)^2}{\left(\sum_i P_i\right)^2 \sum_i (1-T_i)}\right] q_0(1-q_0). \tag{17}$$

Turning now to Ⓓ, and using the fact that $\hat{\gamma}$ is $\{P_k, T_k\}_k-$measurable, we get:

$$\mathbb{E}(\hat{\tau}^{C-IPE}|\{P_k, T_k\}_k) = \hat{\gamma}(\underbrace{\mathbb{E}(\hat{q}_1|\{P_k, T_k\}_k)}_{q_1} - \underbrace{\mathbb{E}(\hat{q}_0|\{P_k, T_k\}_k)}_{q_0}). \tag{18}$$

which gives Ⓓ $= (q_1 - q_0)^2 \mathrm{Var}(\hat{\gamma})$.
Now, using the fact that $\mathrm{Var}(\hat{\gamma}) = \mathbb{E}(\mathrm{Var}(\hat{\gamma}|\{P_k\})) + \mathrm{Var}(\mathbb{E}[\hat{\gamma}|\{P_k\}])$, we get: $\mathrm{Var}(\hat{\gamma}) = \mathbb{E}\left(\frac{1}{\sum_i P_i}\right) \gamma(1-\gamma)$ and finally get

$$Ⓓ = (q_1 - q_0)^2 \mathbb{E}\left[\frac{1}{\sum_i P_i}\right] \gamma(1-\gamma). \tag{19}$$

Combining Equations (15), (17) and (19), we have:

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^{C-IPE}) = &\mathbb{E}\left[\frac{\sum_i T_i}{\left(\sum_i P_i\right)^2}\right] q_1(1-q_1) \\ &+\mathbb{E}\left[\frac{\left(\sum_i T_i\right)^2}{\left(\sum_i P_i\right)^2 \sum_i (1-T_i)}\right] q_0(1-q_0) \\ &+(q_1-q_0)^2\mathbb{E}\left(\frac{1}{\sum_i P_i}\right)\gamma(1-\gamma).\end{aligned} \tag{20}$$

3. Asymptotic variance upper and lower bounds

3.a Asymptotic lower bound of $\mathrm{Var}(\hat{\tau}^{IPE})$

Multiplying (14) by n and using the law of large numbers we get

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\tau}^{IPE}) = 2\left(p_1(1-p_1)+p_0(1-p_0)\right). \tag{21}$$

Now, using $p_1 = (1+\alpha)p_0$, and $\alpha \geq 0$ we can write:

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\tau}^{IPE}) \geq 2p_0(1-p_1)(2+\alpha). \tag{22}$$

3.a Asymptotic upper bound of $\mathrm{Var}(\hat{\tau}^{C-IPE})$

Again, from (20), we have by the law of large numbers

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\tau}^{C-IPE}) = 2\gamma q_1(1-q_1) + \frac{2\gamma^2}{2-\gamma}q_0(1-q_0) + 2\gamma(1-\gamma)(q_1-q_0)^2. \tag{23}$$

Now, using that for any $q \in [0,1]$, $q(1-q) \leq q$, and reminding that $q_1 = (1+\beta)q_0 \leq 1$ where $\beta \geq 0$ by assumption, we get the following asymptotic upper bound for $\mathrm{Var}(\hat{\tau}^{C-IPE})$:

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\tau}^{C-IPE}) \leq 4q_0\gamma(1+\beta). \tag{24}$$

4. Wrap up

We remind that *One-sided non-compliance* and *Exclusive treatment effect* imply straightforwardly that $p_0 = q_0$ (as shown in the beginning of the proof of Proposition 2). Now, combining Equations (22) and (24) (ratio of positive values), we get

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\tau}^{C-IPE})}{\mathrm{Var}(\hat{\tau}^{IPE})} \leq 2\frac{1+\beta}{(1-p_1)(2+\alpha)}\gamma.$$

Taking the square root of this equation delivers the wanted result.

■

5 EXPERIMENTS

To qualify the performance of the C-IPE estimator, we study its benefits in a variety of settings. Firstly we study its properties on simulation-based studies, hereafter denoted by 'Synthetic Datasets', for which (i) the IPE ground truth is known (ii) the level of compliance can be controlled and (iii) we can appreciate performance with respect to an Oracle. Lastly we apply our approach to transform baseline IPE estimators and compare their performance (AUUC) on three real-world, large scale datasets: an open dataset from Criteo, named 'CRITEO-UPLIFT1 Dataset'[1], and two *private* datasets from an online advertising company on which we have privileged access, designated by 'Private datasets'.

[1] `https://ailab.criteo.com/criteo-uplift-prediction-dataset/`

Table 2: Distribution of Users Effectively Treated ($T = 1$) and Visits ($Y = 1$) on CRITEO-UPLIFT1 Split on Treatment Groups

TREATMENT (P)	EXPOSURE (T)	OUTCOME (Y)
0 (2,096,236)	-	3.82% (79,986)
1 (12,161,477)	3.65% (444,384)	4.93% (599,170)

Table 3: Distribution of Visits ($Y = 1$) on CRITEO-UPLIFT1 Projected on Exposure Groups

EXPOSURE (T)	OUTCOME (Y)
0 (13,813,329)	3.58% (495,003)
1 (444,384)	**41.4%** (184,153)

In each experiment we take care of comparing symmetrically formed estimators: we consider baseline treatment effect estimators which we use both as (i) an IPE estimator and as (ii) an ITE estimator to be plugged in the C-IPE decomposition. To simplify experiments we chose two baseline estimators: Two Models (2M) and Shared Data Representation (SDR) as they easily scale to large datasets and have been found competitive in prior studies (Betlei et al., 2018). For reproducibility sake we have implemented all models using the Scikit-Learn Python library (Pedregosa et al., 2011) and all code is available online[2]. All experiments were run on a machine with 48 CPUs (Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz), with 2 Threads per core, and 500Go of RAM. Finally, we note that the state of the art is always evolving and improving. We did not use the most advanced models because we do not aim at outperforming them. Instead, we claim that the C-IPE estimator can improve any IPE estimator if plugged-in as the ITE factor in (7) in a low compliance setting.

5.1 Datasets

Synthetic Datasets We define a simulation setting in which $\mathcal{X} = \{0, 1\}^{10}$, $n = 2.10^6$. The outcome is generated according to:

$$Y \sim \text{Bern}\left(p_0 \left(1 + PT\beta(x)\right)\right), \tag{25}$$

where $P \sim \text{Bern}(0.5)$, $T \sim \text{Bern}(\gamma(x))$, and $p_0 = \mathbb{P}^{\mathfrak{C}}(Y|\overline{T}, x) = 0.1$, using notations from Equations (4). This procedure allows for varying $\gamma(x)$ and $\beta(x)$ to simulate different levels of compliance and relative ITE, respectively.

CRITEO-UPLIFT1 Dataset This open dataset from Criteo contains online advertising data from a randomized controlled trial. Key statistics for this dataset are summarized in Table 2 and 3. Notably, average treatment prescription $\mathbb{E}[P] \approx .85$ indicates that 15% of users were assigned to the control group and shown no advertisement. For the rest of the population, the advertisers participated in online ad auctions. Among the users that advertisers tried to expose ($P = 1$), only 3.65% actually saw an ad, which corresponds to a very low average compliance level $\mathbb{E}[\gamma(X)]$, which we expect to highlight C-IPE estimator benefits. Effective exposure to ads is encoded by the T variable in this setup, while the outcome Y encodes 'at least one visit on the advertiser website'. As illustrated in Figure 2, the mean of Y is more than 10 times higher for exposed users ($T = 1$) than for non-exposed ones ($T = 0$).

[2] `https://github.com/KDD-anonymous-code/individual_treatment_prescription`

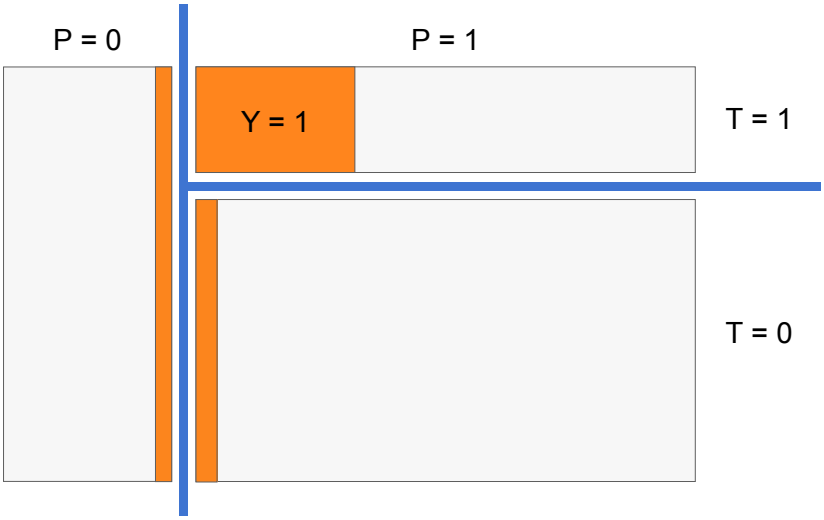

Figure 2: Illustration of the **Exclusive Treatment Effect Assumption** in the CRITEO-UPLIFT1 Dataset: the effect of P on Y goes mainly through the realization of T

Private Datasets We have access to two private datasets of roughly 90 million instances, which contain binary treatment prescription P and outcome Y, and a suitable variable T that embodies an observable compliance. They have a higher average compliance γ (7.8% and 10.4%), higher number of features (48) and similar signal strength than the CRITEO-UPLIFT1 dataset. The purpose of extending the study to these datasets is to verify the relationship between compliance level and performance, predicted by the theoretical study in Section 4.2. The corresponding data has been collected in the same geographical location, during two separate time periods).

5.2 Experiments

In all experiments, nuisance models required by the 2M and SDR estimators are learned as regularized logistic regressions using second order cross-features. This is also the case for the additional nuisance model needed for the C-IPE estimator: the compliance $\hat{\gamma}(.)$. Hyperparameters (regularization norm and strength) of each of these models is carefully selected using internal cross validation, as detailed in appendix.

Compliance Sensitivity Experiment (Simulation) We aim to highlight how the compliance level γ influences the performance of both standard IPE estimators and C-IPE estimators. For this purpose we use a wide range of values $\gamma \in [10^{-4}; .99]$, and generate synthetic datasets as described in Section 5.1 with a different relative ITE values $\beta \in \{-1, 0, 1, 2, 3, 4, 5, 6, 7, 8\}$ for each of the 10 contexts. We report the PEHE metric for both IPE_{2M} and $\text{C} - \text{IPE}_{\text{2M}}$ estimators, and approximate their variance by repeating the experiment with 51 random test/train splits. Recall that PEHE is the mean squared difference between the IPE ground truth and the estimators predictions.

We observe on Figure 3 that the C-IPE estimator significantly outperforms the standard IPE estimator when the level of compliance γ is low, and has similar performance when γ gets closer to 1. This shows that our compliance-aware approach significantly reduces the noise induced by non-compliance and is able to detect smaller IPE signals. This is true in particular for compliance levels γ in the range that is observe on the real datasets that we are using.

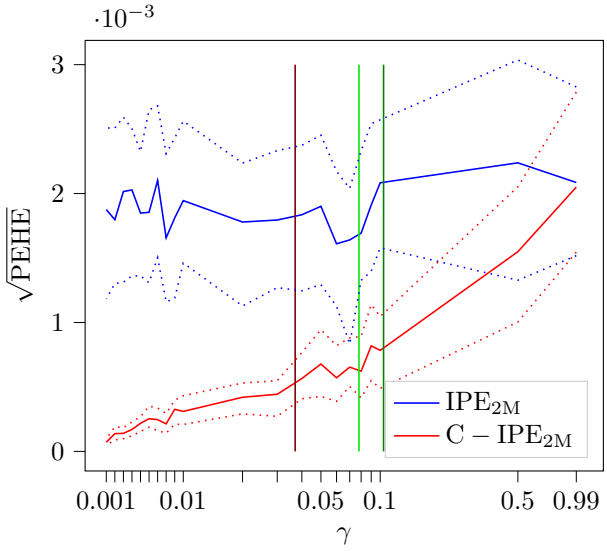

Figure 3: **Compliance Sensitivity (Simulation study)**. PEHE (lower is better) of IPE vs C-IPE estimators at varying compliance level γ. Solid line —— represents the median, and dashed line ····· represents 5% and 95% of confidence intervals. ——, —— and ——: compliance level γ for resp. CRITEO-UPLIFT1 and private datasets 1 & 2. The x-axis has a simlog scale.

Baseline Experiment (Simulation) We simulate a realistic scenario where there exists heterogeneity in compliance γ and treatment effects τ^{IPE}. More precisely, for each instance value x, we draw once and for all $\gamma \in \{0.01, 0.005\}$ and $\beta \in \{-1, 1, 3, 5, 7\}$ uniformly and independently. The associated outcome Y is computed according to (25). We study four methods: standard two-models (IPE_{2M}), shared data representation (IPE_{SDR}) and their compliance-aware variants ($\text{C} - \text{IPE}_{\text{2M}}$, and $\text{C} - \text{IPE}_{\text{SDR}}$). We focus on the AUUC metric, which measures the capacity of the model to rank individuals according to their IPE . The AUUC has the important advantage of not requiring access to the causal effect ground truth, which is typically the case in real-world applications. For more meaningful performance representation, we substract the AUUC of a 'random model' to all raw AUUC values, obtaining ΔAUUC. Finally, for scaling purposes, we also report in Figure 4 results for an Oracle estimator that has access to the drawn (β, γ), and for IPE_{best}, the best possible estimator which does not exploit the observable compliance T (it predicts for each x its empirical IPE average based on the training set). Again, variance is estimated by repeating the experiment with 51 random test/train splits.

Figure 4 assesses the performance of IPE estimators versus their C-IPE variants, using the ΔAUUC metric. C-IPE estimators yield a higher ΔAUUC on more than 90% of the random splits. Moreover the Oracle (best model possible) does not significantly outperform C-IPE estimators, note that even the Oracle can misrank users because the validation set is noisy and empirical IPEs do not always follow the expected ranking. Overall, Figure 4 does not show any limitation of the IPE_{2M} and IPE_{SDR} estimators, but rather highlights the ineffectiveness of such direct IPE estimators in low compliance settings. This phenomenon can be improved by our compliance-aware approach thanks to the higher signal of the causal effect of T on Y. Of course, this synthetic data encodes a simpler setting than real-world data, but the fact that our proposed compliance-aware approach performs that well still confirms our theoretical analysis.

Real-world Experiment (CRITEO-UPLIFT1) To qualify the benefit of C-IPE versus standard IPE for real-world applications we report ΔAUUC on the CRITEO-UPLIFT1 dataset. We study four methods: two IPE estimators (IPE_{2M}, and IPE_{SDR}) and their compliance-aware variants ($\text{C} - \text{IPE}_{\text{2M}}$, and $\text{C} - \text{IPE}_{\text{SDR}}$).

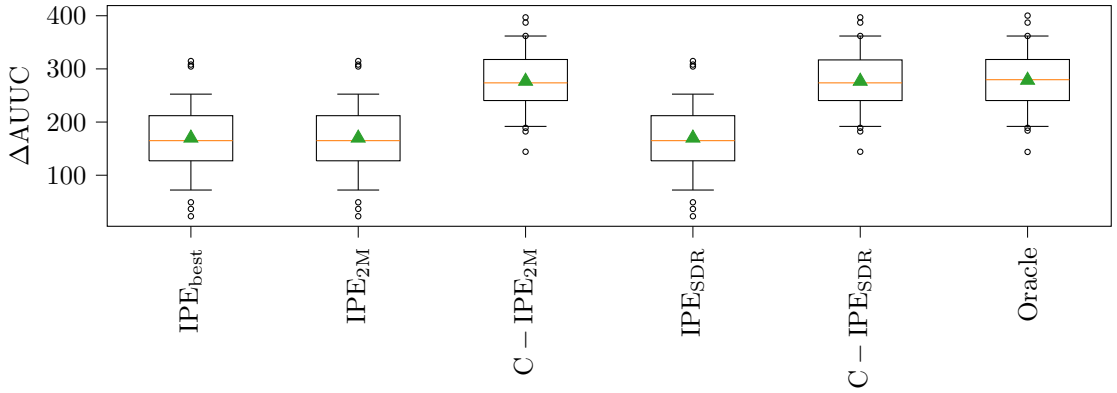

Figure 4: **Baseline Experiment (Simulation)**. ΔAUUC (higher is better) of two IPE estimators, corresponding C-IPE estimators and Oracle model (theoretical truth). Box plots are done on 51 random splits, whiskers at 5/95 percentiles. Note how C-IPE systematically increases AUUC of base estimators.

For the additional $\hat{\gamma}(x)$ model, care is taken to weight the LLH as there is a high imbalance between $T = 1$ and $T = 0$ classes. We use as an hyper-parameter grid the Cartesian product of {L1, L2} (regularization) and $\{0.01, 1, 10^2, 10^5\}$ (C, inverse of regularization strength) for $\hat{\gamma}(x)$. Best hyper-parameters found are L1 regularization and $C = 100$. Results are presented on Figure 5.

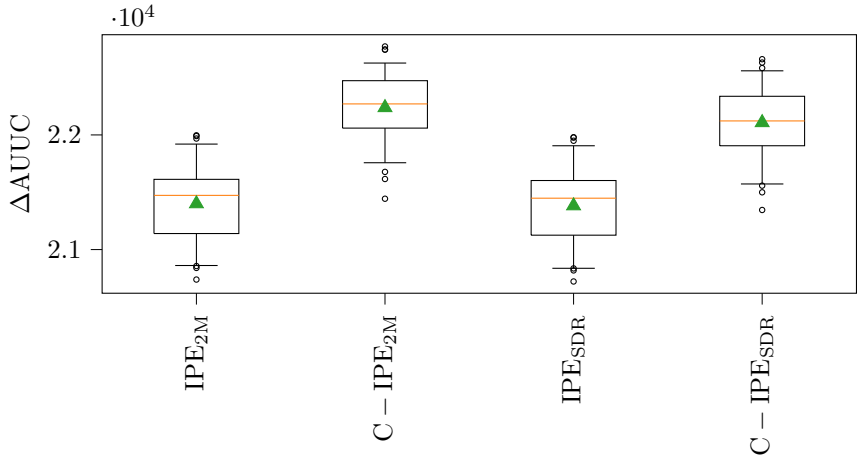

Figure 5: **Real-World Experiment (CRITEO-UPLIFT1)**. ΔAUUC (higher is better) on test data for two IPE estimators and corresponding C-IPE estimators. Box plots computed on 51 random splits, whiskers at 5/95 percentiles. Note the higher ΔAUUC and reduced variance of C-IPE estimators.

The C-IPE version of each estimator has a lower ΔAUUC variance. This was expected from Proposition 3 and Remark 1. Moreover, although the confidence intervals are slightly superposed, C-IPE estimators outperform their IPE counterparts on **each of the 51 splits**.

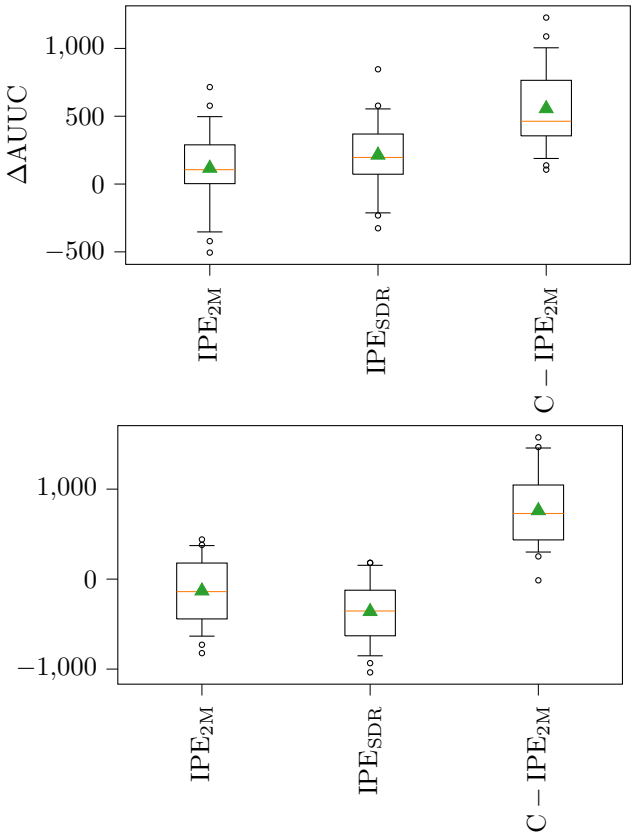

Figure 6: **Real-world Experiment (Private Datasets)**. ΔAUUC (higher is better) of two IPE estimators and one C-IPE estimator (based on two-models). Box plots are done on 33 bootstraps, whiskers at 5/95 percentiles.

Real-world Experiment (Private Datasets)

Figure 6 illustrates the differences in ΔAUUC of the learned models and a random model on 33 bootstraps. The C-IPE models performs better than the two IPE models by having an AUUC significantly better than the random model, when the two IPE models do not perform better than the random model. An interesting finding of this experiment is that, in practice, the mild expected benefit of C-IPE, predicted by the theory (because of a higher compliance than in CRITEO-UPLIFT1 dataset) does not seem discernible. This is however only an indication as there are multiple differences between the two datasets that might explain such a behavior.

6 CONCLUSION AND FUTURE WORK

We propose a novel approach on individual prescription effect (IPE) estimation exploiting observed compliance to the treatment prescription.
Using the structural causal model framework, we define assumptions under which the IPE can be expressed as a product of the individual treatment effect (ITE) and the individual level of compliance. In this setting, our compliance-aware individual prescription effect (C-IPE) estimator is consistent. Moreover its asymptotic variance improves − compared to its standard estimator counterpart − as the level of compliance decreases.

Experimentally, we show how the performance of several baseline IPE estimators improve when plugged in the C-IPE meta-estimator. We also observe that the relationship between estimators performance and compliance behaves as predicted by our theoretical results.
Finally, this work opens several perspectives among which: (i) further study of the stability of our results under variations of assumptions, (ii) derivation of bound tightness guarantees and properties in high-dimensional contexts, and (iii) exploration of how representation learning approaches may uncover by themselves C-IPE-like estimator decomposition under weaker causal assumptions.

References

Ahmed Alaa and Mihaela Schaar. 2018. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*. PMLR, 129–138.

Ahmed Alaa and Mihaela Van Der Schaar. 2019. Validating causal inference models via influence functions. In *Proceedings of the International Conference on Machine Learning*.

Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016).

Susan Athey, Julie Tibshirani, Stefan Wager, et al. 2019. Generalized random forests. *The Annals of Statistics* 47, 2 (2019).

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. 2018. Uplift Prediction with Dependent Feature Representation in Imbalanced Treatment and Control Conditions. In *Proceedings of the International Conference on Neural Information Processing*.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* (2018).

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big data* 6, 1 (2018).

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. 2018. A Large Scale Benchmark for Uplift Modeling. In *Proceedings of the AdKDD and TargetAd Workshop*.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. 2011. Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30, 24 (2011).

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* 38, 2 (2019).

Sandeep K Gupta. 2011. Intention-to-treat concept: a review. *Perspectives in clinical research* 2, 3 (2011), 109.

Miguel A Hernán, James M Robins, et al. 2017. Per-protocol analyses of pragmatic trials. *N Engl J Med* 377, 14 (2017), 1391–1398.

Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

Guido W. Imbens and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* (1994).

Guido W Imbens and Jeffrey M Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47, 1 (2009).

Daniel Jacob, Wolfgang Karl Härdle, and Stefan Lessmann. 2019. Group Average Treatment Effects for Observational Studies. *arXiv preprint arXiv:1911.02688* (2019).

Jing Jin, Grant Edward Sklar, Vernon Min Sen Oh, and Shu Chuen Li. 2008. Factors affecting therapeutic compliance: A review from the patient's perspective. *Therapeutics and clinical risk management* 4, 1 (2008), 269.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 265–274.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019).

Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. 2014. Support vector machines for differential prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. 2015. A cautionary note on the power of the test for the indirect effect in mediation analysis. *Frontiers in Psychology* 5 (2015), 1549.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. 2019. Orthogonal random forest for causal inference. In *Proceedings of the International Conference on Machine Learning*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl. 2017. Detecting latent heterogeneity. *Sociological Methods & Research* 46, 3 (2017).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12 (2011).

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

Nicholas J Radcliffe and Patrick D Surry. 2011. Real-world uplift modelling with significance-based uplift trees. *Stochastic Solutions* (2011).

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005).

Piotr Rzepakowski and Szymon Jaroszewicz. 2010. Decision trees for uplift modeling. In *Proceeding of the International Conference on Data Mining*. IEEE.

Piotr Rzepakowski and Szymon Jaroszewicz. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* 32, 2 (2012).

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. 2019. Machine learning estimation of heterogeneous treatment effects with instruments. In *Proceedings of Advances in Neural Information Processing Systems*.

Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018).

Yu Xie, Jennie E Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological methodology* 42, 1 (2012).

Yuxiang Xie, Nanyu Chen, and Xiaolin Shi. 2018. False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 876–885.

Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. 2018. Uplift modeling from separate labels. In *Proceedings of Advances in Neural Information Processing Systems*.

Xuan Yin and Liangjie Hong. 2019. The Identification and Estimation of Direct and Indirect Effects in A/B Tests through Causal Mediation Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2989–2999.

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

Yao Zhang, Alexis Bellot, and Mihaela Van der Schaar. 2020. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1005–1014.

A PROOFS

A.1 Proof of Proposition 1

First of all, the SCM $\mathfrak{C}$ is Markovian with respect to its own entailed distribution (Proposition 6.31 of Peters et al. (2017)): this implies that every conditional independence encoded in the graph $\mathcal{G}_{\mathfrak{C}}$ holds in distribution $\mathbb{P}^{\mathfrak{C}}$.

Randomized prescription, is implied by the Markov property and the fact that P and X are d-separated by the empty set in $\mathcal{G}_{\mathfrak{C}}$: all paths between P and X contain either $\rightarrow T \leftarrow$ or $\rightarrow Y \leftarrow$, which are blocked by not including neither T nor Y.

Exclusive treatment effect is implied by the Markov property and the fact that the set $\{X, T\}$ d-separates P and Y in $\mathcal{G}_{\mathfrak{C}}$. This is shown by listing all paths between P and Y and observing that they are all blocked by the set $\{X, T\}$.

One-sided non-compliance is straightforwardly implied by the structural assignment of T given in Equations (2).

Valid covariate adjustment relies on the back-door criterion for valid adjustment sets. We remark that $\{X\}$ satisfies the back-door criterion for (T,Y) because (i) it is not a descendant of T and (ii) it blocks all paths from T to Y that enter Y through the backdoor: it is therefore a valid adjustment set for the ordered pair (T, Y).

■

B ADDITIONAL RESULTS AND EXPERIMENTATION DETAILS

B.1 AUUC

AUUC (Rzepakowski and Jaroszewicz, 2012, 2010; Radcliffe and Surry, 2011) is the Area Under the Uplift Curve. It is obtained by ranking the users of a test set according to their predicted uplift, in descending order. In what follows, we will focus on the evaluation of IPEmodels (effect of IPE on Y).

The uplift curve starts at the point $(0, 0)$, then for each user (in decreasing order of the predicted uplift) it goes up of 1 point if the user is in group $P = 1$ with $Y = 1$, it goes down of 1 point if the user is in group $P = 0$ with $Y = 1$, and it stays flat if the user has zero outcome ($Y = 0$). So the label l is $Y * (2P - 1)$. We normalise the x-axis so that it goes from 0 to 1.

The aim of an IPEmodel is to rank first the users with positive outcome and assigned to prescription ($P = 1$), then the users with negative outcome, and finally the users with positive outcome and assigned to non-prescription ($P = 0$). A model respecting such a ranking would have a maximal AUUC.

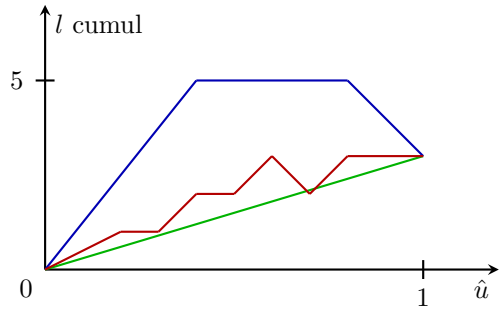

Figure 7: Example of uplift curves for (i) a perfect model (blue), a random model (green) and an example IPE model (red). The AUUC of these models correspond to the area under their respective uplift curves.

The value of the AUUC highly depends on the dataset it is computed on (the test set in the experiments presented in the main text). The first point of the uplift curve is always $(0, 0)$ and the last one is always $(1, U)$ were $U = \sum YP - \sum Y(1 - P)$. The variability of the test set (when we do different splits such as in the experiments presented in the main text) accounts for an important part of the variability of the metric on different splits, which we handle by subtracting the (average) AUUC of a random model − equal to $\frac{U}{2}$ − to the AUUC of the evaluated models: this measure is called ΔAUUC.

B.2 Model training

Our goal is to compare the performance of two approaches for the esimation of the IPE (Individual Treatment Prescription Effect): (1) our proposed C-IPE (compliance-aware Individual Treatment Effect) approach, which exploits observed compliance, and (2) standard IPEestimators which do not exploit observed compliance. We implement two kinds of models for both IPE and ITE (Individual Treatment Intake Effect) factors used inside C-IPEestimators. These models are (1) the Two-Models, with one model learned on the group $P = 1$ and one model learned on the group $P = 0$, and (2) Shared Data Representation (SDR) (Betlei et al., 2018), which is inspired from multi-task learning and has more capacity than the basic Two-Models approach.

On synthetic data, we also compare the models to two theoretical models: IPE_{best} and Oracle:

- IPE_{best} is the best possible model learn-able using training data but without exploiting information from variable T (or equivalently, compliance). In short, this model predicts the difference between the empirical positive outcome rate in the group $P = 1$ and in the group $P = 0$, and does so for each user context $x \in \mathcal{X}$. This approach is valid in the case of our synthetic dataset since we observe a high number of users for each possible context x.
  This is indeed the best possible learnable model, because the dataset is generated with one different probability of positive outcome per category: there is no additional information any model could capture without exploiting the variable T.
  IPE_{best} allows us to show that, in our synthetic dataset, two models and SDR perform close to the best possible IPE approach. In that specific low-dimensional case, there is therefore no need to

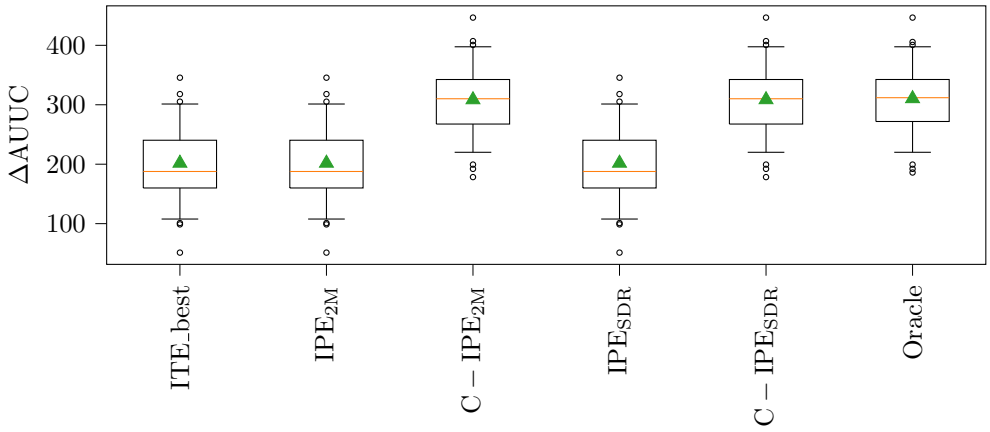

Figure 8: ΔAUUC (higher is better) of two IPE models, corresponding C-IPE models and Oracle model (theoretical truth). Box plots are done on 51 random splits, whiskers at 5/95 percentiles. Note how C-IPE systematically increases the AUUC of standard IPE estimators

implement more complex models (Figure 8) such as doubly-robust methods or tree/forest-based methods.

- Oracle predicts the theoretical ground truth uplift (used to generate the dataset). Its AUUC represents the maximum theoretically reachable AUUC (expected). In practice, we observe that it is however not the best model on all random splits. This can be explained by the fact that the (test) dataset is randomly generated, and that the ranking of users in the test set can therefore differ slightly from the theoretical ranking.

All learned models (IPE, C-IPE and IPE_{best}) suffer from the randomness of the training dataset.

B.3 Model testing

In addition of the randomness of the training dataset, the test dataset is also random. This adds noise to the AUUC values that are computed in practice. We design a metric called, $\text{AUUC}_{\text{thout}}$ that computes the AUUC on a theoretical test set. This metric may only be implemented on synthetic data. It uses the "Theoretical Outcome" (thout) of each category of users as a label, thus circumventing the randomness inherent to the test set generation.

Figure 9 represents the performance of the same models as in Figure 8 but with the $\text{AUUC}_{\text{thout}}$ metric (AUUC on a theoretical test set).

As expected, the Oracle has no variance because it does not − by design − suffer from the randomness of the training set and the $\text{AUUC}_{\text{thout}}$ metric gets rid of the variance of the test set.
C-IPE models also have little to no $\text{AUUC}_{\text{thout}}$ variance in practice. This can be explained by the fact that these models learn treatment intake effect (ITE) signal, which is far less noisy than the IPE signal in low-compliance settings, and therefore suffer from less random variability.

C Code

The full code used for (1) the experiments on synthetic data and (2) experiments on real-world open data described in this paper can be found at `https://github.com/KDD-anonymous-code/individual_treatment_prescription` file, and is formatted so as to **comfortably re-run all experiments**, and compare the results with the ones presented in the paper.

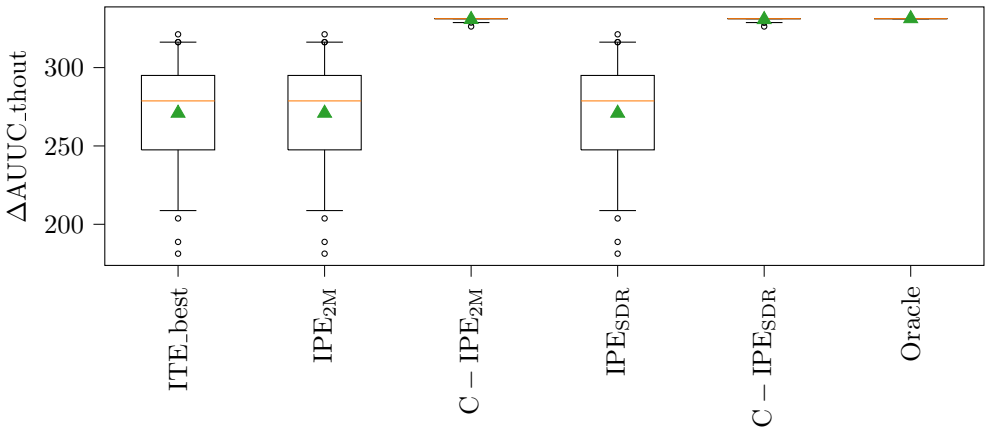

Figure 9: $\Delta AUUC_{thout}$ (higher is better) of two IPE models, corresponding C-IPE models and Oracle model (theoretical truth). Box plots are done on 51 random splits, whiskers at 5/95 percentiles. Note how C-IPE systematically increases the AUUC of standard IPE estimators