Research Article

Wouter A. C. van Amsterdam* and Rajesh Ranganath

Conditional average treatment effect estimation with marginally constrained models

Abstract: Treatment effect estimates are often available from randomized controlled trials as a single *average treatment effect* for a certain patient population. Estimates of the *conditional average treatment effect* (CATE) are more useful for individualized treatment decision-making, but randomized trials are often too small to estimate the CATE. Examples in medical literature make use of the *relative* treatment effect (e.g. an odds ratio) reported by randomized trials to estimate the CATE using large observational datasets. One approach to estimating these CATE models is by using the relative treatment effect as an *offset*, while estimating the covariate-specific untreated risk. We observe that the odds ratios reported in randomized controlled trials are not the odds ratios that are needed in offset models because trials often report the *marginal* odds ratio. We introduce a constraint or a regularizer to better use marginal odds ratios from randomized controlled trials and find that under the standard observational causal inference assumptions, this approach provides a consistent estimate of the CATE. Next, we show that the offset approach is not valid for CATE estimation in the presence of unobserved confounding. We study if the offset assumption and the marginal constraint lead to better approximations of the CATE relative to the alternative of using the average treatment effect estimate from the randomized trial. We empirically show that when the underlying CATE has sufficient variation, the constraint and offset approaches lead to closer approximations to the CATE.

Keywords: conditional average treatment effect, combining observational data and interventional data, causal inference, unobserved confounding

MSC 2020: 62D20, 62F30

1 Introduction

Weighing potential benefits and harms of treatment requires knowing the treatment effect, which is the change in probability of an outcome between different treatments. The gold standard for estimating treatment effects are randomized controlled trials (RCTs). RCTs are designed to provide estimates of the average treatment effect (ATE), which reveals if a treatment has an effect on an outcome. The ATE does not reveal *which patients* would benefit from a treatment. Tailoring the effect to a patient requires knowing how treatment effects change given characteristics of that patient; this is the conditional average treatment effect (CATE). The CATE is a measure of absolute risk, which patients prefer when making decisions [1]. Estimating CATEs directly

---
* Corresponding author: Wouter A. C. van Amsterdam, Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands; Babylon Health Inc., London, UK,
e-mail: w.a.c.vanamsterdam@gmail.com
Rajesh Ranganath: Courant Institute of Mathematical Sciences, Center for Data Science, New York University, New York, United States

from RCT data is often infeasible because trials are generally only powered to estimate population average effects.

Under the standard assumptions like ignorability and positivity, observational data provides an avenue for estimating CATEs [2]. However, techniques for observational causal inference often do not make use of knowing the effect provided by an RCT, which are typically reported on a relative scale as an odds ratio or hazard ratio [3,4]. To make use of population-level relative effects reported by RCT in CATE estimation, several previous studies on breast cancer and cardiovascular disease used the assumption of a *constant relative treatment effect* to develop CATE models from observational data [5–8]. We call this assumption the constant-relative ATE (CR-ATE) assumption and models that use this assumption CR-ATE models.

The assumption of a constant relative treatment effect does not imply that the CATE must be constant because even with a constant relative treatment effect, the treatment can have a varying effect on an absolute risk scale depending on the untreated risk of a patient. For instance, assume that a new cholesterol lowering drug reduces the risk of cardiovascular death within the next 10 years with an odds ratio of 0.5. A 60-year-old male smoker with hypertension and increased cholesterol has an untreated risk of cardiovascular death of 40% and should expect a reduction in risk of 15% points. A 50-year-old female without hypertension has an untreated risk of under 1% and will have a less than 0.5% points reduction in risk. Given these widely different effects on an absolute probability scale, one may recommend the new cholesterol lowering drug to the 60-year-old male but not the 50-year-old female.

One approach to use the CR-ATE assumption with observational data is to use the constant relative treatment effect as an *offset* term in a model. Some models built with an offset term were found to be accurate in observational validation studies, on the basis of which treatment guidelines acknowledged a place for them in clinical decision-making [9,10]. Because CR-ATE models target interventional distributions, the use of such model implicitly assumes that the RCT's relative effect provides the correct offset and that its use controls for unobserved confounding. However, whether the assumption is correct has not been discussed or verified.

In this work, we evaluate the validity of the assumption that a known constant odds ratio for treatment allows for CATE estimation when using the known odds ratio as an offset term. Under the standard requirements for observational causal inference, we show that there is a mismatch in that the odds ratios reported in RCT are not the odds ratios that are needed for the offset method because RCT generally report estimates of the *marginal* odds ratio, whereas the offset method requires the *conditional* odds ratio. Further, when there is unobserved confounding, we demonstrate that even with the correct conditional odds ratio offsets are not sufficient for estimating CATEs.

To address the mismatch in odds ratios, we introduce a new estimator called the marginally constrained model (MCM). The MCM constrains a model's implied marginal odds ratio to be close to that reported from an RCT. We show that with no unobserved confounding, this approach is a consistent estimator of the CATE, and we empirically show a large reduction in the number of samples needed to estimate the CATE when using this marginal odds ratio constraint. We also show how this regularizer can be combined with the CR-ATE assumption in constant-relative marginally constrained model (CR-MCM) models.

Next we turn to violations of ignorability, i.e. the assumption of no unobserved confounding. Without ignorability in offset models, consistent estimation of CATEs is not possible. However, the goal here is to study if models like the CR-MCM approximate the CATE better than the alternative provided by randomized trials, the ATE. The measure of better is the precision in heterogeneous treatment effect estimation (PEHE) [11]. For example, the ATE is a good approximation of the CATE when the underlying CATE has little variation as a function of the conditioning set. Better approximations to the CATE can lead to patient benefit as the CATE leads to better individualized treatment decisions. We empirically show in the absence of ignorability that MCMs almost always lead to better CATE estimation than the average treatment effect-baseline (ATE-baseline) whenever there is sufficient variation in the underlying CATE. Finally, we demonstrate the utility of the CR-ATE assumption and show that when the assumption holds, CR-MCMs are even better than MCMs.

2 Offset models: estimation and non-collapsibility

We consider models for the absolute difference in probability of a binary outcome Y under two possible treatments $T \in \{0, 1\}$ conditional on a pre-treatment covariate vector X, using data from the observational distribution $q(Y, T, X)$. The estimation target is the CATE, conditional on X. Without loss of generality, treatment $T = 0$ is assumed to be the baseline treatment (or no treatment depending on the clinical context) and $T = 1$ is the comparator treatment of interest. We denote Y_t as the potential outcome of Y if T is set to t by intervention, and $Y_t|x$ as the analogous potential outcome conditional on $X = x$. We refer to $\Pr(Y_0 = 1|X)$ as the *untreated risk*, meaning the probability of experiencing the outcome when assigned no treatment (or the control treatment), conditional on X. The CATE is defined as follows:

$$\text{CATE}(x) := \Pr(Y_1 = 1|x) - \Pr(Y_0 = 1|x) \tag{1}$$

Taking the expected value over X of the CATE, we obtain the ATE. Working backward from the definition of the ATE:

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y_1 - Y_0] \\ &= \Pr(Y_1 = 1) - \Pr(Y_0 = 1) \\ &= \mathbb{E}[\Pr(Y_1 = 1|x)] - \mathbb{E}[\Pr(Y_0 = 1|x)] \\ &= \mathbb{E}[\Pr(Y_1 = 1|x) - \Pr(Y_0 = 1|x)] \\ &= \mathbb{E}[\text{CATE}(x)]. \end{aligned} \tag{2}$$

The ATE (equation (2)) is *one* measure of treatment effect based on the *marginal* potential outcomes Y_0, Y_1. In addition to the ATE expressed in absolute probabilities, a common measure of treatment effect for binary outcomes employed by RCT is the *marginal odds ratio* [3,4,12–14]:

$$\text{OR}(Y_1, Y_0) = \frac{\Pr(Y_1 = 1)(1 - \Pr(Y_0 = 1))}{(1 - \Pr(Y_1 = 1))\Pr(Y_0 = 1)}. \tag{3}$$

It is often more convenient to work with the log odds ratio. Writing $\sigma(x) = (1 + e^{-x})^{-1}$ as the *sigmoid* function (also known as the *logistic* function) with its inverse σ^{-1} (known as the *logit* function), the log odds ratio denoted as γ is given as follows:

$$\gamma = \gamma(Y_1, Y_0) = \log \text{OR}(Y_1, Y_0) = \sigma^{-1}(\Pr(Y_1 = 1)) - \sigma^{-1}(\Pr(Y_0 = 1)). \tag{4}$$

The log odds ratio γ as a function of potential outcomes Y_0, Y_1 is a measure of causal treatment effect. Throughout we will assume to have access to γ from a RCT conducted in the same population as the observational study. We assume that this RCT provides an unbiased estimate of the treatment effect measure γ with infinite precision. However, as is typically the case due to data-sharing restrictions, only γ is available from the RCT and not the original data. Although RCT can estimate many other parameters of $Y_t|X$ such as the conditional odds ratio [15,16], in practice, they often only report a marginal effect. In Section 6, we describe ways to relax these assumptions.

2.1 Offset models as CATE models

We now study of a class of CR-ATE models that are already used in clinical practice: *offset models* [5]. Relying on an assumed functional form of $\Pr(Y_t|X)$, offset models aim to approximate $\Pr(Y_t|X)$ in a constrained way by forcing the "effect" of treatment to be equal to some known, constant, *relative treatment effect* on an appropriate scale. In the context of binary outcomes, one CR-ATE assumption is that the odds ratio for treatment is constant for all X, whereas the *untreated* risk $\Pr(Y_0 = 1|X)$ varies with X, meaning that:

$$\Pr(Y_t = 1|x) = \sigma(\sigma^{-1}(\Pr(Y_0 = 1|x)) + \beta_t t). \tag{5}$$

The assumption in equation (5) is called the *offset assumption* for logistic models and is an example of a constant-relative treatment effect assumption. Note that β_t is a measure of treatment effect derived from the

conditional potential outcomes $Y_0|X$, $Y_1|X$, and the specific offset assumption is that β_t does not depend on X. Using this assumption with also assuming β_t is known yields a class of models of $\Pr(Y_t = 1|X)$ called $g : \{0, 1\} \times \mathcal{X} \to [0, 1]$, defined as follows:

$$g(t, x) = \sigma(\sigma^{-1}(g_0(x)) + \beta_t t), \tag{6}$$

where $g_0 = g(0, x) : \mathcal{X} \to [0, 1]$. In the context of generalized linear models, a fixed term in a model that is not estimated from data is called an *offset* term [17]. We therefore refer to models of the form of equation (6) as *treatment offset models* or *offset models* for short. Later we will study under what conditions offset models lead to consistent estimators of the CATE.

Offset models and analogous CR-ATE models have been used in practice without any justification [5–8]. Although the parametric assumption in equation (5) may seem strong, one supporting argument is that the odds ratio was a more stable measure of the treatment effect than the absolute risk difference in a review of 125 meta-analyses of RCT [18]. In addition, RCT rarely find evidence for variation in the odds ratio depending on covariates (i.e. interaction terms between covariates and treatment on the log-odds scale), though it should be noted that RCT is generally underpowered for estimating these interaction terms. Either way, the fact that offset models are used in practice warrants a formal understanding of their consistency and under what conditions they are likely to provide a better approximation to the CATE than the ATE-baseline.

2.2 Estimation of offset models

To study offset models, we must first describe how they may be estimated. One can construct an estimator for offset models of the form in equation (6) by specifying a class of functions $\mathcal{G}_0 = \{g_0 \in \mathcal{G}_0 : \mathcal{X} \to [0, 1]\}$ for the untreated risk $\Pr(Y_0 = 1|X)$. Assume that X is discrete with cardinality d. Coding X as a d-dimensional one-hot vector leads to a natural parameterization for non-parametric models of $\Pr(Y_0 = 1|X)$ with d logistic regression parameters $\boldsymbol{b} \in \mathbb{R}^d$, giving rise to model family $\mathcal{G}_0 = \{g_0 : \mathcal{X} \to [0, 1], g_0(x;\boldsymbol{b}) = \sigma(\boldsymbol{b}'x), \boldsymbol{b} \in \mathbb{R}^d\}$. This $\mathcal{G}_0$ together with a known β_t defines a family of offset models:

$$\mathcal{G} = \{g : \{0, 1\} \times \mathcal{X} \to [0, 1], g(t, x;\boldsymbol{b}, \beta_t) = \sigma(\boldsymbol{b}'x + \beta_t t), \boldsymbol{b} \in \mathbb{R}^d\}. \tag{7}$$

Parameter vector $\boldsymbol{b}$ may be obtained by maximizing the likelihood of the observed data. The question is under what conditions this yields a consistent estimator of $\Pr(Y_t = 1|X)$. If for every $x \in \mathcal{X}$, $0 < \Pr(Y_0 = 1|X = x) < 1$, there is always a $\boldsymbol{b}^* \in \mathbb{R}^d$ such that $\Pr(Y_0 = 1|x) = \sigma(\boldsymbol{b}^{*\prime}x)$. With this $\boldsymbol{b}^*$, we have that by the offset assumption:

$$g^*(t, x; \boldsymbol{b}^*) := \sigma(\boldsymbol{b}^{*\prime}x + \beta_t t) = \sigma(\sigma^{-1}(\Pr(Y_0 = 1|x)) + \beta_t t) = \Pr(Y_t = 1|x).$$

2.3 Collapsibility

An important concept when considering offset models is *collapsibility*. A measure $\mu(Y_0, Y_1)$ of causal effect is said to be *collapsible* over variable X if there exists a set of weights w_x such that $\mu(Y_0, Y_1) = \frac{1}{\sum w_x}\sum w_x \mu(Y_0, Y_1|x)$, meaning that the marginal effect is a weighted average of the x-conditional effects [19,20]. As shown in equation (2), the ATE is a collapsible effect measure by weighting the CATEs by the probability of X. For odds ratios, except for very special circumstances, there are no such weights meaning that the odds ratio is *non-collapsible* [21,22].

The non-collapsibility of the odds ratio creates a problem for estimating logistic offset models as the *marginal log odds ratio γ* reported in the RCT is not equal to the *conditional log odds ratio β_t* required in equation (6). This means that the model in equation (6) cannot be estimated from the available data. The stronger the T-conditional association between X and Y, the greater the difference between γ and β_t [23]. For an illustration, see Appendix A.1. This mismatch is an important drawback as at the same time, a stronger association between X and Y conditional on T results in more variation in $\Pr(Y_0|X)$ and thus more variation in the CATE under the constant-relative treatment effect assumption. So in the situation where offset models have

more potential added value (when the CATEs vary substantially), the estimate of the marginal log odds ratio γ from RCT becomes a less accurate approximation of the conditional odds ratio β_t needed for estimating the offset model. If one were to use γ in the place of β_t in equation (6), this would lead to an inconsistent estimator. We provide a numerical example in Section 5.2 that highlights this effect of non-collapsibility on offset models.

3 Marginally constrained models

We now turn to a new class of CATE models that also exploits knowledge from prior RCT. There are many instances where an estimate of γ is available from the published results of an RCT, but not the CATE because the sample size of the RCT was too small. Running a new, bigger RCT may be infeasible due to costs, or deemed unethical because of the absence of equipoise, i.e. uncertainty about what treatment is superior if the prior RCT demonstrated that one treatment was superior over the other *on average*. When turning to observational data to estimate the CATEs, instead of ignoring the γ estimate from the RCT, we can use it in the estimation as a constraint. We describe a procedure for incorporating γ in CATE estimation based on the marginalization of predicted outcome probabilities. Under some regularity conditions, we prove that this method leads to a consistent estimator of the CATE when the standard causal inference assumptions hold.

3.1 Exploiting RCT evidence by using the marginal odds ratio as a constraint

Assume we have a function g from a family of functions $\mathcal{G} = \{g : \{0, 1\} \times X \to [0, 1]\}$ where we interpret $g(t, x)$ as a predicted probability $\Pr(Y = 1 | t, x) = g(t, x)$. Given a distribution q over X, we can calculate the marginalized predictions of g: $\Pr(Y = 1 | t) = \mathbb{E}_{x \sim q(x)} g(t, x)$. Correspondingly we have the *implied marginal log odds ratio* $M(g)$ (implied by g and q, but suppressing the dependency on q in the notation) by using the marginalized predictions of g over q.

$$M(g) = \sigma^{-1}(\mathbb{E}g(1, X)) - \sigma^{-1}(\mathbb{E}g(0, X)). \tag{8}$$

Given an independent and identically distributed sample of X with sample size n, define the empirical counterpart of M as follows:

$$M_n(g) = \sigma^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(1, x_i)\right) - \sigma^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(0, x_i)\right). \tag{9}$$

This empirical marginalizer was used in ref. [22]. Assume g was found by maximizing the log-likelihood of the observed data over function class $\mathcal{G}$, $L_n : \mathcal{G} \to \mathbb{R}$. We can augment the optimization objective using the known marginal odds ratio γ from the prior RCT as follows:

$$\mathcal{L}_n(g) = L_n(g) - \lambda(M_n(g) - \gamma)^2, \lambda > 0. \tag{10}$$

We call optimization with the objective a marginally constrained model. If $\mathcal{G}$ encompasses all conditional distributions on $\{0, 1\} \times X$, i.e. in the non-parametric estimation setting, it follows that $\exists g* \in \mathcal{G}$ such that $g^*(t, x) = \Pr(Y_t = 1 | x)$. Assume that additionally the standard observational causal inference assumptions hold, namely, $(Y_1, Y_0) \perp\!\!\!\perp T | X$ (strong ignorability), $0 < q(T | X) < 1$ (positivity) and $Y_t = Y$ if $T = t$ (consistency). For more details on these assumptions, see e.g. [24,25]. Theorem 1 states that under these conditions, the augmented objective 10 yields a consistent estimator of $\Pr(Y_t = 1 | X)$.

Theorem (informal) 1. *Given binary treatment T, binary outcome Y and covariate X. Assume strong ignorability $(Y_1, Y_0) \perp\!\!\!\perp T | X$, positivity $0 < q(T | X) < 1$ and consistency $Y_t = Y$ if $T = t$. Given a family of functions $\mathcal{G} = \{g \in \mathcal{G} : \{0, 1\} \times X \to [0, 1]\}$ and assuming $\exists g^* \in \mathcal{G}, \Pr(Y_t = 1 | x) = g^*(t, x)$. Denote the sample log-likelihood $L_n : \mathcal{G} \to \mathbb{R}$. In addition, given marginal log odds ratio $\gamma = \sigma^{-1}(\Pr(Y_1 = 1)) - \sigma^{-1}(\Pr(Y_0 = 1))$, sample marginalizer $M_n : \mathcal{G} \to \mathbb{R}$ and $\lambda > 0$. Then:*

$$g_n = \arg\max_{g \in \mathcal{G}}[L_n(g) - \lambda(M_n(g) - \gamma)^2] \text{ is a consistent estimator of } \Pr(Y_t = 1 | X).$$

For proving consistency, additional assumptions on convergence and identifiability of the observational objective are required. The formal theorem statement and proof are provided in the Appendix A.2. Under the conditions of Theorem 1, both the unconstrained maximum-likelihood estimator and the marginally constrained estimator are consistent estimators of the CATE. However, as the MCM optimizes over a smaller family of functions, we should expect it to be more statistically efficient. We test the relative efficiency of the MCM versus the unconstrained estimator in Section 5.1.

3.2 Marginally constrained constant-relative treatment effect model estimation

MCM leverages knowledge of the *marginal* odds ratio γ for CATE estimation, but it does not use the offset-assumption from equation (5). Offset models do rely on this assumption but often cannot be estimated because the required *conditional* odds ratio is not known. We now introduce the CR-MCM, a CATE model that leverages both the offset assumption and a known *marginal* odds ratio γ. With discrete X of cardinality d, non-parametric modelling of $\Pr(Y_t|X)$ requires $2d$ parameters, d parameters for $\Pr(Y_0|X)$ and d additional parameters for $\Pr(Y_1|X)$. The offset assumption in equation (5) implies that there exists a $\boldsymbol{b}^* \in \mathbb{R}^d$ and β_t such that $\sigma(\boldsymbol{b}^{*'}x + \beta_t t) = \Pr(Y_t = 1|x)$. To exploit the offset assumption in MCMs and reduce the number of parameters, we can use the constrained objective in equation (10) and consider optimizing over the offset model family in equation (7). However, without access to β_t, the model family is underspecified, and we cannot proceed with optimization. Instead we introduce CR-MCM as follows:

$$\boldsymbol{b}, b_t = \arg\max_{\boldsymbol{b}\in\mathbb{R}^d, b_t\in\mathbb{R}} [L_n(g(t, x; \boldsymbol{b}, b_t)) - \lambda(M_n(g(t, x; \boldsymbol{b}, b_t)) - \gamma)^2], \lambda > 0 \tag{11}$$

The model family for CR-MCM has $d + 1$ parameters and by the offset assumption $\exists \boldsymbol{b}^*, b_t^* \in \mathbb{R}^d \times \mathbb{R}$, $g(t, x; \boldsymbol{b}^*, b_t^*) = \Pr(Y_t = 1|x)$. Because this model family is a correctly specified generalized linear model, maximizing the (unconstrained) likelihood over this family is a consistent estimator of $\Pr(Y = 1|T, X)$. Given strong ignorability, positivity and consistency, $\Pr(Y = 1|T, X) = \Pr(Y_t = 1|X)$. Under the regularity conditions of Theorem 1, we have that equation (11) is a also consistent estimator of $\Pr(Y|T, X)$. Moreover, CR-MCM may be more efficient than MCM when the offset assumption holds.

4 Offset models and MCMs for CATE approximation under unobserved confounding

We now turn to the case where the standard assumption of strong ignorability does not hold: in the presence of unobserved confounding. In the presence of unobserved confounding, meaning that Y_t is not independent of T conditional on X, CATE estimators based on adjustment cannot be used. However, this does not preclude causal effect estimation as there are known settings where causal effects can be determined despite the presence of unobserved confounding if one can rely on additional assumptions regarding the data generating mechanism. Example methods are instrumental variable estimation [26–28] and methods based on proxy-variables of unmeasured confounders [29,30].

4.1 Offset models under unobserved confounding

We investigate whether the structural assumption made in the offset model in equation (5) combined with knowledge of the conditional odds ratio β_t makes the offset model a consistent estimator of the CATE. With a simple counter example, we prove that this is not the case.

4.1.1 Example 1: Offset models are inconsistent estimators in the presence of unobserved confounding

A simple example compatible with the offset assumption in equation (5) is when there is a binary unobserved confounder U but no covariate X. Denoting $\mathcal{B}$ as the Bernoulli distribution and $q_u = \Pr(U = 1)$, then the data-generating mechanism for this example is:

$$u \sim \mathcal{B}(q_u), t \sim \mathcal{B}(\Pr(T = 1|u)), y \sim \mathcal{B}(\Pr(Y_t = 1|u)). \tag{12}$$

Given the offset model family from equation (7), a natural parameterization of $g(t, x) = g(t)$ in the context of this example is $g(t;b_0) = \sigma(b_0 + \beta_t t)$, $b_0 \in \mathbb{R}$. To disentangle the issue of non-collapsibility from unobserved confounding, we assume that β_t is given *a priori* and is not estimated. We derive a closed form expression for the expected log-likelihood of the observational data depending on the single parameter b_0 of the offset model $L(b_0)$ in Appendix A.3. By taking the derivative with respect to b_0 and plugging in the ground truth value $\beta_0 \coloneqq \sigma^{-1}(\Pr(Y_0 = 1))$, we find the following expression:

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0) = q_u(1 - q_u)[(\Pr(Y|T = 0, U = 1) - \Pr(Y|T = 0, U = 0))(q(T = 0|U = 1) - q(T = 0|U = 0))$$
$$+ (\Pr(Y|T = 1, U = 1) - \Pr(Y|T = 1, U = 0))(q(T = 1|U = 1) - q(T = 1|U = 0))].$$

In general, this expression is non-zero, meaning that the ground truth solution β_0 is not a stationary point of the expected log-likelihood. Thereby, the offset method is an inconsistent estimator for $\Pr(Y_0 = 1)$ in the presence of unobserved confounding. When either $q(T|U = 1) = q(T|U = 0)$ or $\Pr(Y_t = 1|U = 1) = \Pr(Y_t = 1|U = 0)$, the derivative is zero at β_0, meaning that β_0 is a stationary point of the expected log-likelihood. Either of these cases imply no unobserved confounding.

Despite its simplicity, this example is important for all offset models with discrete X because: (a) when the treatment is binary, any arbitrary unobserved confounder can be modelled as a single binary variable while maintaining the same observational and interventional distributions [31]; and (b) given X with cardinality $d > 1$, optimizing over the offset model family in equation (7) is equivalent to stratifying the population for each value of X and optimizing the same objective as in Example 1 in each stratum. Thus, if the offset model is an inconsistent approximator of $\Pr(Y_0 = 1)$ in this simple example with no covariate X, optimization over the offset model family for discrete X will also be inconsistent for $\Pr(Y_0|X)$ by implication, even when the correct conditional log odds ratio β_t is known. In Appendix A.3, we show that this implies that the offset model is inconsistent for the CATE as well, aside from very rare chance occasions.

4.2 Metric and ATE baseline

Offset models are not consistent estimators of the CATE in the presence of unobserved confounding and require knowledge of the conditional odds ratio which is generally not available. However, offset models may still be useful for treatment decisions if they lead to better CATE approximation than what is used in current clinical practice. As RCT generally only estimate a single ATE, this is often what current treatment decisions are based on. Ideally, treatment decisions are based on the difference in probability of the outcome under different treatments conditional on patient characteristics, i.e. the CATE. A direct measure for how well a model approximates the CATE is the root-mean-square error of approximated versus actual CATE. In the context of CATE approximation, this metric is sometimes called the PEHE [11]. The PEHE for function $g : \{0, 1\} \times \mathcal{X} \rightarrow [0, 1]$ is defined as follows:

$$\text{PEHE}(g) = \sqrt{\mathbb{E}(\text{CATE}(x) - (g(1, x) - g(0, x)))^2} \tag{13}$$

As RCTs only provide the ATE, in our experiments, we use the PEHE of the ATE as the baseline. If the CATE is not constant but varies with X, the ATE is a bad approximator of the CATE, leading to suboptimal PEHE and thus suboptimal treatment decisions. Measuring the approximation error with the PEHE is how we arrive at what we call the ATE-baseline: the PEHE that is obtained in the current situation by using a single treatment effect for treatment decisions in all patients, instead of using the CATE. The PEHE of the ATE-baseline is calculated as follows:

$$\text{PEHE}(\text{ATE}) = \sqrt{\mathbb{E}[(\text{CATE}(x) - \text{ATE})^2]}. \tag{14}$$

Given that $\text{ATE} = \mathbb{E}[\text{CATE}(x)]$, the ATE-baseline has an intuitive form:

$$
\begin{aligned}
\text{PEHE}(\text{ATE}) &= \sqrt{\mathbb{E}[(\text{CATE}(x) - \text{ATE})^2]} \\
&= \sqrt{\mathbb{E}[(\text{CATE}(x) - \mathbb{E}[\text{CATE}(x)])^2]} \\
&= \sqrt{\text{VAR}(\text{CATE})} \\
&= \text{SD}(\text{CATE}).
\end{aligned} \tag{15}
$$

For CATE approximation models such as the offset model and the MCM and CR-MCM, even if the model is inconsistent under certain conditions, it may still be a valid modelling choice if it has lower PEHE than the ATE baseline because it can lead to better treatment decisions compared with current clinical practice.

5 Experiments

We now study the different models for CATE approximation in four experiments. In Appendix A.4, we derive a closed form solution for the parameters of the offset model in the case of a binary covariate. This closed form can be used to characterize the bias in offset solutions, though the resulting formula is opaque. Therefore, we use the closed form for an extensive empirical evaluation. First, we investigate the relative efficiency of the MCM compared to unconstrained estimation when both are consistent. Then we investigate the PEHE of the offset model under (a) non-collapsibility but no unobserved confounding and (b) unobserved confounding but no non-collapsibility. Finally, we compare the PEHE of three different CATE approximators, the offset model with marginal odds ratio γ, MCM and CR-MCM, and study when these models have better PEHE than the ATE-baseline in a large experimental grid with a binary covariate and an unobserved confounder U.

Implementation. For implementing MCM and CR-MCM, we optimize the constrained objective 10 by specifying a small value $\lambda_0 = 0.01$ and optimizing the resulting unconstrained objective with a limited-memory Broyden-Fletcher-Goldfarb-Shanno optimizer. After convergence of the unconstrained optimizer yielding model estimate $\hat{g}$, we evaluate the constraint $(M(\hat{g}) - \gamma)^2 < \varepsilon = 10^{-4}$. If this is not satisfied, for the next iteration $i = 1, 2, \ldots$, we increase λ such that $\lambda_i = 10 * \lambda_{i-1}$ and repeat until the constraint is satisfied. We relied on the JAX [32], JAXOpt [33] and NumPyro [34] python libraries for implementing the experiments. The code to reproduce all experiments is publicly available at www.github.com/vanamsterdam/binarymcm (DOI: 10.5281/zenodo.8144896).

5.1 Relative efficiency of the marginally constrained estimator

Under the assumptions of strong ignorability, positivity and consistency and with a known marginal odds ratio γ, we have at least two consistent estimators of the CATE: the unconstrained maximum-likelihood estimator and the MCM. To test the relative efficiency of the constrained versus the unconstrained estimator, we define an experimental grid with a binary covariate X and data generating mechanism $\sigma^{-1}(\text{Pr}(Y_t = 1|x)) = \beta_0 + \beta_t t + \beta_x x$. The parameter grid is given in Table 1.

In this setup, a non-parametric model family for $\text{Pr}(Y_t = 1|X)$ is, given as follows:

$$\mathcal{G} = \{g : \{0, 1\}^2 \to [0, 1], g(t, x) = \sigma(b_0 + b_t t + b_x x + b_{tx} tx), b_0, b_t, b_x, b_{tx} \in \mathbb{R}\}.$$

Given this model family, sample log-likelihood estimator L_n, sample marginalizer M_n and marginal log odds ratio γ (calculated from the parameters of the data generating mechanism), we compare the constrained and unconstrained estimator. For each of the possible combinations of parameter values in Table 1, we generated 250 datasets and applied both estimators. Each fitted model produces an estimate of $\text{CATE}(X = 0)$ and $\text{CATE}(X = 1)$. We construct 95% confidence bounds for these estimated CATEs by taking the 2.5 and 97.5 percentile values over the 250 repetitions for each simulation setting, sample size, estimator and X. The width of each confidence bound (len(CI)) is a measure of statistical uncertainty that is relevant for treatment decision-

Table 1: Grid for experiment on relative efficiency of the marginally constrained versus the unconstrained estimator

Parameter	Definition	Values
q_x	$\Pr(X = 1)$	0.5
β_0	$\sigma^{-1}(\Pr(Y_0 = 1|X = 0))$	$\sigma^{-1}(0.15, 0.5)$
β_t	$\sigma^{-1}(\Pr(Y_1 = 1|X))-\sigma^{-1}(\Pr(Y_0 = 1|X))$	1
β_x	$\sigma^{-1}(\Pr(Y_t = 1|X = 1))-\sigma^{-1}(\Pr(Y_t = 1|X = 0))$	log(1/5, 1/2, 1, 2, 5)
q_t	$\Pr(T = 1)$	0.5
η_x	$\sigma^{-1}(\Pr(T = 1|X = 1))-\sigma^{-1}(\Pr(T = 1|X = 0))$	log(1/5, 1/2, 1, 2, 5)
n	Sample size	50, 75, 100, 150, 200, 325, 400, 500, 750

making. As expected by the asymptotic theory for asymptotically linear normal estimators, we found empirically that $\log(\text{len(CI)})$ was approximately linear in the logarithm of the sample size for each parameter setting and estimator, see Figures A2 and A3 in the Appendix. We fit models to the experimental results to summarize the relative efficiency. Specifically, by denoting $m = 1$ to indicate the unconstrained estimator and $m = 0$ for the constrained estimator, we fit linear regression models of the form $\log n = w_0 + w_l \log(\text{len(CI)}) + w_m m + \varepsilon$ to the experimental results of each parameter combination. We took len(CI) to be the average of the logarithm of the confidence bound length of CATE($X = 1$) and CATE($X = 0$) in each experiment.

The linear models provided a good fit of the experimental results with an adjusted $R^2 > 0.975$ for all of the combinations of simulation parameters. The parameter w_m in this linear model has the following interpretation: if sample size n_c reaches a certain length of the confidence bound with the constrained estimator, to reach the same length of the confidence bound with the unconstrained method, $n_c e^{w_m}$ samples are needed, meaning a $100(e^{w_m} − 1)\%$ increase in the sample size. Across all simulation settings, we find that unconstrained estimation requires at least 71% and at most 106% more samples, meaning that to achieve equal-width confidence bounds in the CATE estimation, one needs at least 71% more patients with the unconstrained estimator then when using the constrained estimator MCM.
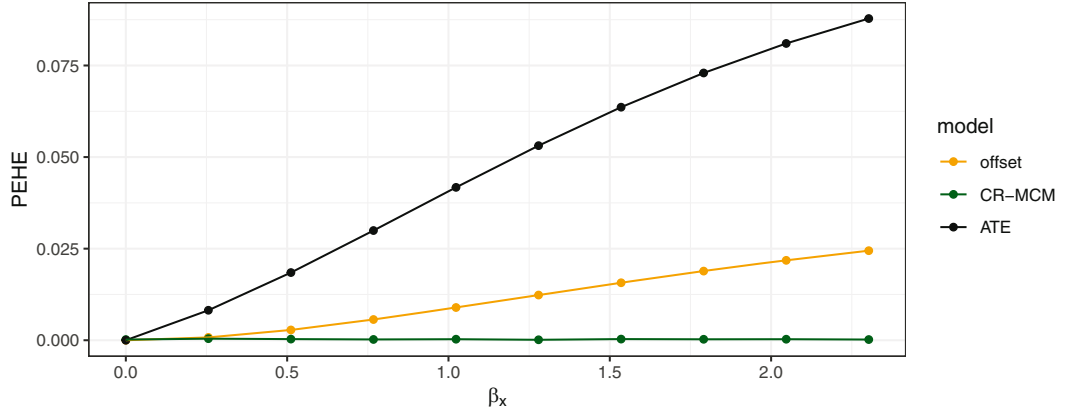
5.2 The effect of non-collapsibility on the offset method

When the offset method is applied using the *marginal* log odds ratio γ in the place of *conditional* log odds ratio β_t, the offset model is inconsistent and has non-zero PEHE. We conducted a numerical experiment to evaluate the PEHE of the offset model compared with the ATE-baseline and CR-MCM with a single binary covariate X with varying effects on Y. The data-generating mechanism for this experiment is $\sigma^{-1}(\Pr(Y_t = 1|x)) = \beta_0 + \beta_t t + \beta_x x$, with $\beta_0 = 0, \beta_t = 1, 0 \le \beta_x \le \log 15 = 2.71$. Furthermore, $\Pr(X = 1) = \Pr(T = 1) = 0.5$.

In Figure 1, we see that when β_x increases, the PEHE of the offset model with γ as an offset increases as expected due to the increasing difference between γ and β_t. However, the PEHE of the ATE-baseline increases faster. Finally, we note that the CR-MCM remains consistent as by Theorem 1.

5.3 The effect of unobserved confounding on the offset method

When the strong ignorability does not hold, for example due to the presence of an unobserved confounder U, Example 1 (Section 4.1.1) demonstrates that the offset model is an inconsistent estimator of $\Pr(Y_t = 1)$, even if the conditional log odds ratio β_t is known. As motivated earlier, it may still be justifiable to use the offset method when it has better PEHE than the ATE-baseline. To study the effect of unobserved confounding on the PEHE of the offset method, we conducted an experiment with a binary covariate X, binary unobserved confounder U and in this case a known *conditional* log odds ratio β_t. We note that this β_t is generally not

Figure 1: Experiment on the effect of non-collapsibility on the offset model when the *marginal* log odds ratio γ is used instead of the *conditional* log odds ratio β_t, for varying effects of X on Y. When the outcome depends more on covariate x (i.e. bigger β_x), the difference between the *marginal* odds ratio and *conditional* odds ratio becomes bigger due to non-collapsibility of the odds ratio, leading to worse CATE approximation error for the offset model. The CR-MCM model uses the marginal odds ratio as a constraint and remains consistent. CR-MCM: constant-relative marginally constrained model, ATE: average treatment effect.

available from RCTs, but we use it in this experiment to separate out the effect of unobserved confounding from non-collapsibility. The data-generating mechanism for this experiment is:

- $\sigma^{-1}(\Pr(Y_t = 1 | X = x, U = u)) = \alpha_0 + \alpha_t t + \alpha_x x + \alpha_{tx} tx + \alpha_u u$
- $\Pr(U | T = t)$ such that $\sigma^{-1}(\Pr(U = 1 | T = 1)) - \sigma^{-1}(\Pr(U = 1 | T = 0)) = \gamma_u = \alpha_u$.

The parameter ranges were $\alpha_0 = \sigma^{-1}(0.05)$, $\alpha_t = 1$, $\alpha_x \in \log(1, 2, 3, 4, 5)$, and $\alpha_u \in \log(1, 2, 5)$. We set the marginals $\Pr(U = 1) = \Pr(X = 1) = \Pr(T = 1) = 0.5$. Parameter α_x controls the effect of X on Y and α_u determines the amount of unobserved confounding (both through $U \rightarrow Y$ and $U \rightarrow T$). Note that if the offset assumption in equation (5) is used, it is assumed to hold conditional on the observed covariate X, i.e. for distribution $\Pr(Y_t | X)$, not for $\Pr(Y_t | X, U)$. Therefore, given values for $\alpha_0, \alpha_t, \alpha_x, \alpha_u$, and $\Pr(U | T = t)$, the parameter α_{tx} was determined such that the offset assumption was valid for $\Pr(Y_t | X)$. To implement CR-MCM, the *marginal* log odds ratio for treatment γ was calculated from the experiment parameters.

On this experimental grid, four different models are compared: the ATE-baseline, the offset method with conditional odds ratio, CR-MCM and a "fully-observational" estimator, i.e. the non-parametric maximum likelihood estimator of $q(Y = 1 | T, X)$, where q denotes the observational distribution. The results are presented in Figure 2. As expected, the PEHE of the fully observational estimator is very sensitive to the increasing amounts of unobserved confounding. The PEHE of the ATE-baseline increases when the variance of the CATE increases, which in this experiment is determined by α_x. While the PEHE for the offset method increases when unobserved confounding increases, it still has better PEHE than the ATE-baseline when $\alpha_x \neq 0$. Finally, CR-MCM is the best performing model overall even though, in contrast with the offset model, does not require knowledge of the conditional odds ratio β_t, which is generally not available from RCTs.

5.4 Marginally constrained CATE approximation in the presence of unobserved confounding and non-collapsibility

Finally, we investigate whether different CATE approximators lead to better PEHE than the ATE-baseline in an experimental grid covering the space of configurations for a binary covariate X and binary unobserved confounder U. The parameter grid and values are presented in Tables 2 and 3.

For the parameter α_{tx}, we took one of two approaches depending on "enforce offset": either one of the values listed in Table 3 when "enforce offset" = False, or the value that makes the offset assumption satisfied in $\Pr(Y_t = 1 | X)$ when "enforce offset" = True.
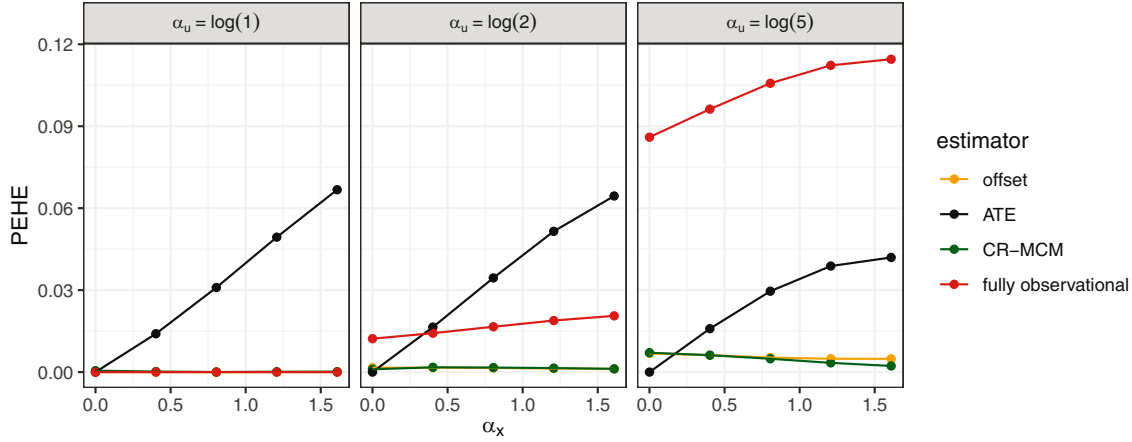
Figure 2: Experiment on the effect of unobserved confounding on the offset estimator when the *conditional* log odds ratio β_t is known, but there is an unobserved confounder, for varying effects of X on Y. α_x determines the effect of X on the outcome Y and α_u determines the amount of unobserved confounding. When α_x obtains bigger, the ATE becomes a bad approximator for the CATE. When unobserved confounding increases, the fully observational baseline becomes a bad approximator of the CATE as expected. Both the offset model and the CR-MCM perform well, although the offset model requires knowledge of the conditional odds ratio which is generally not available, while CR-MCM only requires the marginal odds ratio which is generally reported in RCTs. CR-MCM: constant-relative marginally constrained model, ATE: average treatment effect.

To summarize the results, we formulated a criterion based on the variance of the CATE as when the CATE is constant, the PEHE of the ATE-baseline is 0 and no model can improve on that. The higher the variance of the CATE, the more CATE models can potentially improve on the ATE-baseline. We group experiments by setting a threshold τ for the variance of the CATE. To parameterize τ, we turn to the scale of the CATEs, under the assumption that $\Pr(X = 1) = 0.5$, the simplest version of the experiment. This makes τ a function of only the CATEs, which makes it more interpretable. To express τ on the CATE scale, we introduce δ as the difference between the CATEs depending on X:

$$\delta := |\mathrm{CATE}(X = 1) - \mathrm{CATE}(X = 0)|$$

Given δ and $\Pr(X = 1) = 0.5$, we can calculate τ as follows:

$$
\begin{aligned}
\tau &= \mathrm{VAR}(\mathrm{CATE}) \\
&= \mathbb{E}_x[(\mathrm{CATE}(x) - \mathbb{E}[\mathrm{CATE}(x)])^2] \\
&= \Pr(X = 0)(\mathrm{CATE}(X = 0) - (\Pr(X = 0)\mathrm{CATE}(X = 0) + \Pr(X = 1)\mathrm{CATE}(X = 1)))^2 \\
&\quad + \Pr(X = 1)(\mathrm{CATE}(X = 1) - (\Pr(X = 0)\mathrm{CATE}(X = 0) + \Pr(X = 1)\mathrm{CATE}(X = 1)))^2 \\
&= 0.5(0.5\,\mathrm{CATE}(X = 0) - 0.5\,\mathrm{CATE}(X = 1))^2 + 0.5(0.5\,\mathrm{CATE}(X = 1) - 0.5\,\mathrm{CATE}(X = 0))^2 \\
&= 0.5 * 0.25 * (\delta^2 + \delta^2) \\
&= 0.25\delta^2.
\end{aligned}
$$

For each τ, we subsetted the experiments for which $\mathrm{VAR}(\mathrm{CATE}) > \tau$ and we evaluated in what percentage of experiments the CATE approximators performed better than the ATE-baseline. Furthermore, we split the results for whether the offset assumption did or did not hold.

Table 2: Grid for experiment on PEHE of CATE approximators versus ATE-baseline, definition of parameters

Distribution	Parameters
$\Pr(Y_t\|X)$	$\Pr(Y = 1\|t, x, u) = \sigma(\alpha_0 + \alpha_t t + \alpha_x x + \alpha_u u + \alpha_{tx} tx + \alpha_{tu} tu + \alpha_{xu} xu + \alpha_{txu} txu)$
$\Pr(U\|T)$	$\gamma_u = \sigma^{-1}(\Pr(U = 1\|T = 1)) - \sigma^{-1}(\Pr(U = 1\|T = 0))$
$\Pr(U)$	$\Pr(U = 1)$
$\Pr(X)$	$\Pr(X = 1)$
$\Pr(T)$	$\Pr(T = 1)$

Table 3: Grid for experiment on PEHE of CR-MCM versus ATE-baseline, parameter values

Values	Parameters
logit(0.025, 0.15, 0.5, 0.85)	α_0
0, 0.402, 0.804, 0.121, 1.161	$\alpha_x, \alpha_u, \alpha_{tu}, \alpha_{xu}, \alpha_{txu}, \alpha_{tx}^\dagger$
0.402, 0.804, 0.121, 1.161	α_t
–1.161, –1.207, –0.804, –0.402, 0, 0.402, 0.804, 1.207, 1.161	γ_u
0.2, 0.5	$\Pr(U = 1), \Pr(X = 1), \Pr(T = 1)$
True, False	Enforce offset

$^\dagger \alpha_{tx}$ takes these values only when enforce offset = False.

We took $\delta \in \{0.01, 0.05\}$ corresponding to relatively small differences in CATEs. As shown in Table 4, when δ increases, for all CATE approximators, the fraction of experiments with improved PEHE compared to the ATE-baseline increases. The offset model with marginal odds ratio and the CR-MCM improve relative to MCM when the offset assumption is indeed satisfied. Taking the experiments in which the offset assumption is satisfied together with those in which it is not, MCM performs best with >91% of experiments having better PEHE than the ATE-baseline when $\delta > 0$, >97% when $\delta > 0.01$ and >99% when $\delta > 0.05$. Even though the offset model with marginal odds ratio is the worst performing model in the comparison, it still leads to better PEHE estimation than the ATE-baseline when $\delta > 0.05$ in >80% of the experiments when the offset assumption does not hold, and >98% of experiments where the offset assumption does hold. This indicates that the CR-ATE models in clinical use [5,6] might provide value for individualized treatment decision-making even if they were developed in data with unobserved confounding.

6 Discussion

We introduced MCMs that make use of a known marginal treatment effect to approximate the CATE from observational data. Under the standard causal inference assumptions and some regularity conditions, MCMs are consistent estimators of the CATE, and in our experiments, they are also more efficient than unconstrained estimation. Next, in the presence of unobserved confounding, we showed that the offset method does not provide consistent CATE estimates for binary outcomes. We find that MCMs tend to have better PEHE than offset models, and both models have better PEHE than the ATE-baseline in almost all settings in the case of a binary covariate, as long as the variance in the CATE is greater than a minimal threshold. Further, when the constant relative effect assumption holds, CR-MCMs are even better.

Table 4: PEHE of different CATE approximators for varying values of $\delta \doteq |\text{CATE}(X = 1) - \text{CATE}(X = 0)|$

δ	Offset satisfied	Offset	CR-MCM	MCM	N
0.00	False	0.68	0.70	**0.96**	16,455,472
0.00	True	0.70	**0.77**	0.69	3,282,598
0.01	False	0.71	0.73	**0.98**	15,777,278
0.01	True	0.90	**0.97**	0.89	2,497,022
0.05	False	0.80	0.81	**1.00**	13,193,906
0.05	True	0.98	**0.99**	0.98	1,295,294

The experiments were filtered such that $V\,AR(\text{CATE}) > \frac{1}{4}\delta^2$, which is the variance of the CATE for that value of δ when $\Pr(X = 1) = 0.5$. The numbers in columns offset, CR-MCM and MCM denote in what percentage of the experiments that model had better PEHE than the ATE baseline. offset: offset approach with *marginal* odds ratio used in the place of the conditional odds ratio, CR-MCM: constant-relative marginally constrained model, MCM: marginally constrained model, N: number of experiments satisfying the VAR(CATE) and "offset satisfied" requirement.

An important question for the offset model and the CR-MCM is when it is valid to assume that the relative treatment effect is indeed constant conditional on the patient features. There is some evidence from meta-analyses that treatment effect estimates on a relative scale are more stable across different RCTs than treatment effects on an absolute scale [18]. However, in some settings, there may be clear indications for differences in treatment effect on a relative scale. For instance, breast cancer patients respond better to estrogen receptor modulator tamoxifen when they have an estrogen-sensitive tumour [35]. When the difference in relative treatment effect is known, this difference could be accounted for accordingly in MCM and offset models. The constant odds ratio for treatment remains a strong parametric assumption, though in our experiments we found that the offset model and CR-MCM tend to have better PEHE than the ATE-baseline even if the offset assumption does not hold, as long as the variance in the CATE is greater than a minimal threshold.

Instead of only using a single treatment effect estimate from prior RCTs, recent work has studied combining observational data and data from RCTs for CATE estimation [31,36]. Under relatively mild assumptions, estimates from combined datasets yield more efficient estimates of CATEs than using RCT data alone. However, these methods require access to the individual patient data from the RCT, whereas MCMs only rely on a single-effect estimate from RCTs, which is usually available from the published RCT results. Gaining access to individual patient data from RCTs is often challenging due to data-access restrictions.

We only studied settings with a single binary covariate in our experiments. Future work could experiment with higher dimensional, mixed-type covariates. In higher dimensions, the constraint on the implied marginal odds ratio restricts a lower fraction of the degrees of freedom. It is unknown whether the constraint will help attain better CATE approximation in the presence of unobserved confounding with higher dimensional parameter spaces, and the relative efficiency gain of using the constraint may be smaller. Machine learning methods that learn lower-dimensional representations of covariate distributions that still preserve the information relevant for the untreated risk might help restore the efficiency.

Future work could extend our experiments to relative treatment effect estimates for time-to-event outcomes such as the hazard ratio or to the settings of time-varying treatments and confounding. Furthermore, we assumed that the RCT gives an unbiased estimate of γ with infinite precision. In practice, RCTs are often conducted in non-random samples of the population, which may result in different covariate distributions $p_{\mathrm{rct}}(X) \neq q_{\mathrm{obs}}(X)$. If $\Pr(Y_t|X)$ is transportable from the RCT to the observational data, the marginal odds ratio from the RCT will be different from the implied marginal odds ratio of $\Pr(Y_t|X)$ when calculated in the observational distribution, because $\Pr(Y_t|X)$ is marginalized over a different distribution of X. However, if $p_{\mathrm{rct}}(X)$ and $q_{\mathrm{obs}}(X)$ are known and appropriate sampling weights exist, the constraint on the marginal odds ratio may be applied using these weights.

Bayesian extensions of MCMs can be investigated to account for uncertainty in marginal odds ratio estimates from RCTs. Finally, finite-sample characteristics of the estimator for the implied marginal odds ratio in terms of bias and variance could be studied further.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The code to reproduce all experiments is publicly available at www.github.com/vanamsterdam/binarymcm (DOI: 10.5281/zenodo.8144896).

References

[1] Murray EJ, Caniglia EC, Swanson SA, Hernández-Diiiaz S, Hernán MA. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. J Clin Epidemiol. 2018 Nov;103:10–21.

[2] Causal diagrams and the identification of causal effects. In: Pearl J, editor. Causality. Cambridge: Cambridge University Press; 2009. p. 65–106. https://www.cambridge.org/core/books/causality/causal-diagrams-and-the-identification-of-causal-effects/ D9AE074727C3AC9AFE9F0CD4C7A506B5.

[3] Furie R, Rovin BH, Houssiau F, Malvar A, Teng YKO, Contreras G, et al. Two-year, randomized, controlled trial of Belimumab in Lupus Nephritis. New England J Med. 2020 Sep;383(12):1117–28.

[4] Lean ME, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. Lancet (London, England). 2018 Feb;391(10120):541–51.

[5] Candido dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res. 2017 Dec;19(1):58, 80 citations (Crossref) [2021-08-06]. http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-017-0852-3.

[6] Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol. 2001 Feb;19(4):980–91. 679 citations (Crossref) [2021-08-06]. http://ascopubs.org/doi/10.1200/JCO.2001.19.4.980.

[7] Alaa AM, Gurdasani D, Harris AL, Rashbass J, van der Schaar M. Machine learning to guide the use of adjuvant therapies for breast cancer. Nature Machine Intelligence. 2021 Aug;3(8):716–26. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Breast cancer; Prognosis Subject_term_id: breast-cancer;prognosis. https://www.nature.com/articles/s42256-021-00353-8.

[8] Xu Z, Arnold M, Stevens D, Kaptoge S, Pennells L, Sweeting MJ, et al. Prediction of cardiovascular disease risk accounting for future initiation of statin treatment. Am J Epidemiol. 2021;190(10):2000–14. https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab031/6140872.

[9] Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019 Aug;30(8):1194–220. https://linkinghub.elsevier.com/retrieve/pii/S0923753419312876.

[10] Gradishar WJ. NCCN Breast Cancer Guideline, Version 5.2021; 2021. https://www.nccn.org/professionals/physician_gls/pdf/breast-2.pdf.

[11] Hill JL. Bayesian nonparametric modelling for causal inference. J Comput Graph Stat. 2011 Jan;20(1):217–40. http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.08162.

[12] Suverein MM, Delnoij TSR, Lorusso R, Brandon Bravo Bruinsma GJ, Otterspoor L, Elzo Kraemer CV, et al. Early extracorporeal CPR for refractory out-of-hospital cardiac arrest. New England J Med. 2023 Jan;388(4):299–309. http://www.nejm.org/doi/10.1056/NEJMoa2204511.

[13] Hill KD, Kannankeril PJ, Jacobs JP, Baldwin HS, Jacobs ML, O'Brien SM, et al. Methylprednisolone for heart surgery in infants - A randomized, controlled trial. New England J Med. 2022 Dec;387(23):2138–49. http://www.nejm.org/doi/10.1056/NEJMoa2212667.

[14] Weghuber D, Barrett T, Barrientos-Pérez M, Gies I, Hesse D, Jeppesen OK, et al. Once-weekly semaglutide in adolescents with obesity. New England J Med. 2022 Dec;387(24):2245–57. http://www.nejm.org/doi/10.1056/NEJMoa2208601.

[15] Tchetgen Tchetgen EJ, Robins JM, Rotnitzky A. On doubly robust estimation in a semiparametric odds ratio model. Biometrika. 2010 Mar;97(1):171–80. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412601/.

[16] Tchetgen Tchetgen EJ, Rotnitzky A. Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. Stat Med. 2011 Feb;30(4):335–47.

[17] Watson T. Practitioneras guide to generalized linear models. Arlington Virginia, US: Towers Watson; 2007.

[18] Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000 Jul;19(13):1707–28.

[19] Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. Emerging Themes Epidemiol. 2019 Jan;16(1):1. doi: 10.1186/s12982-018-0083-9.

[20] Didelez V, Stensrud MJ. On the logic of collapsibility for causal effect measures. Biometr J. 2022;64(2):235–42. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.202000305. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202000305.

[21] Whittemore AS. Collapsibility of multidimensional contingency tables. J R Stat Soc Ser B (Methodol). 1978;40(3):328–40. Publisher: [Royal Statistical Society, Wiley]. https://www.jstor.org/stable/2984697.

[22] Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biometr J. 2021;63(3):528–57. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201900297. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201900297.

[23] Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol. 1991 Jan;44(1):77–81. https://www.sciencedirect.com/science/article/pii/089543569190203L.

[24] Pearl J. Causal inference in statistics: An overview. Statistics Surveys. 2009 Jan;3(none). https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.full.

[25]   Hernan MA, Robins JM. Causal Inference: What If. 2020.

[26]   Wald A. The fitting of straight lines if both variables are subject to error. Ann Math Stat. 1940 Sep;11(3):284–300. Publisher: Institute of Mathematical Statistics. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-11/issue-3/The-Fitting-of-Straight-Lines-if-Both-Variables-are-Subject/10.1214/aoms/1177731868.full.

[27]   Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep IV: A flexible approach for counterfactual prediction. In: International Conference on Machine Learning. PMLR; 2017. p. 1414–23.

[28]   Puli A, Ranganath R. General control functions for causal effect estimation from IVs. Adv Neural Inform Process Syst. 2020;33:8440–51.

[29]   Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika. 2018 Dec;105(4):987–93. doi: 10.1093/biomet/asy038.

[30]   van Amsterdam WAC, Verhoeff JJC, Harlianto NI, Bartholomeus GA, Puli AM, de Jong PA, et al. Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer. Scientific Reports. 2022 Apr;12(1):5848. Number: 1 Publisher: Nature Publishing Group. https://www.nature.com/articles/s41598-022-09775-9.

[31]   Ilse M, Forré P, Welling M, Mooij JM. Combining interventional and observational data using causal reductions. 2022 Jan. http://arxiv.org/abs/2103.04786.

[32]   Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al. JAX: composable transformations of Python+NumPy programs; 2018. http://github.com/google/jax.

[33]   Blondel M, Berthet Q, Cuturi M, Frostig R, Hoyer S, Llinares-López F, et al. Efficient and modular implicit differentiation. 2022. [cs, math, stat]. http://arxiv.org/abs/2105.15183.

[34]   Phan D, Pradhan N, Jankowiak M. Composable effects for flexible and accelerated probabilistic programming in NumPyro. 2019. arXiv: http://arXiv.org/abs/arXiv:191211554.

[35]   Group EBCTC. Tamoxifen for early breast cancer: an overview of the randomised trials. The Lancet. 1998 May;351(9114):1451–67. https://www.sciencedirect.com/science/article/pii/S0140673697114234.

[36]   Rosenman E, Basse G, Owen A, Baiocchi M. Combining observational and experimental datasets using shrinkage estimators. [math, stat]. 2020 May. http://arxiv.org/abs/2002.06708.

Appendix

A.1 Non-collapsibility

Here, we provide an example and intution on what non-collapsibility of the odds ratio is and why the difference between the conditional odds ratio and the marginal odds ratio increases when the assocation between x and y becomes greater. Consider the following data-generating mechanism for binary X with $q(X = 1) = 0.5$, binary treatment T, and outcome mechanism $\Pr(Y_t = 1|x) = \sigma(b_0(x) + t)$, so that the *conditional* odds ratio ($e^1 \approx 2.72$) is constant. As we will see, depending on how b_0 depends on x, the *marginal* log odds ratio γ_t will vary. For two settings of $b_0(x)$, we calculate the resulting *marginal* odds ratio γ_t in a few simple steps. The calculations are visualized in Figure A1. Let $\pi_t(x) = \Pr(Y_t = 1|x)$:

$$\pi_0(0) = \sigma(b_0(x = 0))$$
$$\pi_0(1) = \sigma(b_0(x = 1))$$
$$\pi_1(0) = \sigma(b_0(x = 0) + 1)$$
$$\pi_1(1) = \sigma(b_0(x = 1) + 1)$$
$$\pi_0 = (1 - q(x = 1))\pi_0(0) + q(x = 1)\pi_0(1)$$
$$\pi_1 = (1 - q(x = 1))\pi_1(0) + q(x = 1)\pi_1(1)$$
$$\eta_0 = \sigma^{-1}(\pi_0)$$
$$\eta_1 = \sigma^{-1}(\pi_1)$$
$$\gamma_t = \eta_1 - \eta_0.$$
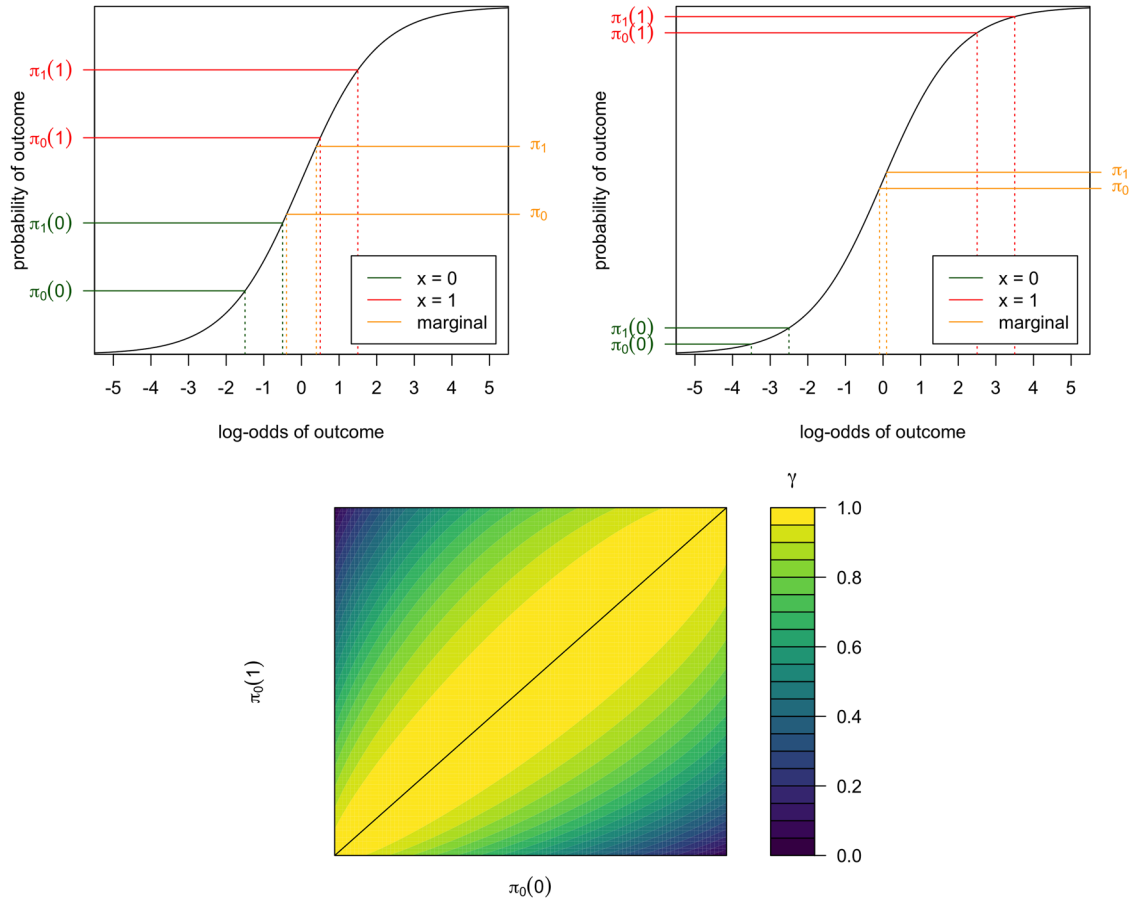
Figure A1: Illustration of non-collapsibility. For fixed $p(x = 1) = 0.5$ and $\beta_t = 1.0$, the marginal log odds ratio γ of treatment becomes closer to 0 when the difference in the untreated risks $\pi_0(0)$, $\pi_0(1)$ becomes larger.

Table A1: Numeric illustration of non-collapsibility of the odds-ratio. The marginal odds-ratio γ is closer to zero than the conditional odds-ratio β_t

Setting	x	$\eta_0(x)$	$\eta_1(x)$	β_t	$\pi_0(x)$	$\pi_1(x)$	π_0	π_1	η_0	η_1	γ_t
a	0	1.5	0.5	1	0.182	0.378	0.402	0.598	0.395	0.395	0.791
	1	0.5	1.5	1	0.622	0.818					
b	0	3.5	2.5	1	0.029	0.076	0.477	0.523	0.093	0.093	0.186
	1	2.5	3.5	1	0.924	0.971					

This leads to the following numerical results in Table A1 where we see that $\beta_t > \gamma_t > 0$ and $\gamma_t \to 0$ when the difference between $\pi_0(0)$, $\pi_0(1)$ becomes bigger, despite $\beta_t = 1$ remaining constant.

A.2 Consistency of marginally constrained models

Here, we prove Theorem 1. First, we state the formal version of the theorem:

Theorem A1. (Formal version) *Given binary treatment T, binary outcome Y and covariate X. Assume strong ignorability $(Y_1, Y_0) \perp\!\!\!\perp T|X$, positivity $0 < q(T|X) < 1$ and consistency $Y_t = Y$ if $T = t$. Given a model $p_\theta(y|t, x)$ indexed by parameter $\theta \in \Theta$, and assuming $\exists \theta^* \in \Theta$, $\Pr(Y_t = 1|X = x) = p_{\theta^*}(y|t, x)$. Denote the sample log-likelihood $L_n : \Theta \to \mathbb{R}$, marginal log odds ratio $\gamma = \sigma^{-1}(\Pr(Y_1 = 1)) - \sigma^{-1}(\Pr(Y_0 = 1))$, sample marginalizer $M_n : \Theta \to \mathbb{R}$:*

$$M_n(\theta) = \sigma^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} p_\theta(Y = 1|1, x_i)\right) - \sigma^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} p_\theta(Y = 1|0, x_i)\right) \tag{A1}$$

and $0 < \lambda < \infty$. Assume that L_n converges uniformly in probability to $L : \Theta \to \mathbb{R}$ and M_n to $M : \Theta \to \mathbb{R}$:

$$M(\theta) = \sigma^{-1}(\mathbb{E}\,p_\theta(Y = 1|1, X)) - \sigma^{-1}(\mathbb{E}\,p_\theta(Y = 1|0, X)) \tag{A2}$$

The uniform convergence in probability means:

$$\sup_{\theta \in \Theta}|L_n(\theta) - L(\theta)| \overset{p}{\to} 0$$

$$\sup_{\theta \in \Theta}|M_n(\theta) - M(\theta)| \overset{p}{\to} 0.$$

In addition, assume strong identifiability, namely, that for every $\varepsilon > 0$, the Kullback-Leibler (KL) divergence between the data distribution and model distribution:

$$\inf_{\theta:|\theta - \theta^*| \geq \varepsilon} \mathrm{KL}\left(p_{\theta^*}(y|t, x), p_\theta(y|t, x)\right) > 0$$

Then:

$$\theta_n = \arg\max_{\theta \in \Theta}[L_n(\theta) - \lambda(M_n(\theta) - \gamma)^2]$$

is a consistent estimator of θ^* so the model matches $\Pr(Y_t = 1|X)$.

To prove this theorem, we first prove that $\Pr(Y_t = 1|X) = \Pr(Y|T, X)$ (causal identifiability) and then consistency of the constrained estimator for the observational distribution.

Proof. A standard causal inference result is that under strong ignorability, positivity and (causal) consistency, $\Pr(Y_t = 1|X) = \Pr(Y = 1|T, X)$. Restating the proof:

$$\Pr(Y_t = 1|X) = \Pr(Y_t = 1|X, t) = \Pr(Y = 1|X, t).$$

The first equality follows from the strong ingorability assumption $(Y_0, Y_1) \perp\!\!\!\perp T|X$, and the second follows from the consistency assumption $Y_t = Y$, $T = t$. The right-hand side now only contains observable quantities. We now prove that the constrained estimator is a consistent estimator of $\Pr(Y_t = 1|X)$.

Writing the population objective and its empirical counterpart:

$$\mathcal{L}(\theta) = \mathbb{E} \log p_\theta(y_i|x_i, t_i) - \lambda(M(\theta) - \gamma)^2$$
$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i|x_i, t_i) - \lambda(M_n(\theta) - \gamma)^2. \tag{A3}$$

We first introduce a new objective:

$$R_n^{\theta^*}(\theta) = \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^*}(y_i|x_i, t_i)}{p_\theta(y_i|x_i, t_i)} - \lambda(M_n(\theta^*) - \gamma)^2 + \lambda(M_n(\theta) - \gamma)^2. \tag{A4}$$

Because the first term in A4 is constant, it is clear that

$$\hat{\theta} = \arg\min_{\theta \in \Theta} R_n^{\theta^*}(\theta) \Leftrightarrow \hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

The population version of $R_n^{\theta^*}$ is:

$$\begin{aligned}
R^{\theta^*}(\theta) &= \mathbb{E}_{t,x}\mathbb{E}_{y|t,x} \log \frac{p_{\theta^*}(y|x, t)}{p_\theta(y|x, t)} - \lambda(M(\theta^*) - \gamma)^2 + \lambda(M(\theta) - \gamma)^2 \\
&\overset{(i)}{=} \mathbb{E}_{t,x}\mathrm{KL}(p_{\theta^*}(y|x, t)||p_\theta(y|x, t)) - \lambda(M(\theta^*) - \gamma)^2 + \lambda(M(\theta) - \gamma)^2 \\
&\overset{(ii)}{=} \mathbb{E}_{t,x}\mathrm{KL}(p_{\theta^*}(y|x, t)||p_\theta(y|x, t)) + \lambda(M(\theta) - \gamma)^2,
\end{aligned} \tag{A5}$$

where (i) is substituting the definition of the KL-divergence, and equality (ii) holds as $\gamma = M(\theta^*)$ by definition.

Lemma A3 gives that uniform convergence in probability of L_n and M_n imply uniform convergence in probability of $\mathcal{L}_n$ and thus R_n as R_n is just a constant minus $\mathcal{L}_n$. We now use this convergence to prove that $R_n^{\theta^*} \overset{p}{\to} 0$ as $n \to \infty$, which will prove the consistency of the constrained estimator. Denote $\hat{\theta}_n$ as the maximizer of $\mathcal{L}_n$, because of this, it must be that

$$R_n^{\theta^*}(\hat{\theta}_n) = \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\hat{\theta}_n) \leq 0.$$

We can now bound $R^{\theta^*}(\hat{\theta}_n)$ and show that it converges to 0 in probability:

$$R^{\theta^*}(\hat{\theta}_n) = R^{\theta^*}(\hat{\theta}_n) - R_n^{\theta^*}(\hat{\theta}_n) + R_n^{\theta^*}(\hat{\theta}_n) \overset{(i)}{\leq} R^{\theta^*}(\hat{\theta}_n) - R_n^{\theta^*}(\hat{\theta}_n) \overset{p}{\to} 0. \tag{A6}$$

The inequality (i) holds as $R_n^\theta(\hat{\theta}) \leq 0$ and the convergence is given by uniform convergence of R_n. To show that the KL-divergence goes to zero as well, note that:

$$R^{\theta^*}(\hat{\theta}_n) = \underbrace{\mathbb{E}_{t,x}\mathrm{KL}(p_{\theta^*}(y|x, t)||p_{\hat{\theta}_n}(y|x, t))}_{A_n} + \underbrace{\lambda(M(\hat{\theta}_n) - \gamma)^2}_{B_n} \overset{p}{\to} 0. \tag{A7}$$

A_n is non-negative because the KL divergence is non-negative and B_n is non-negative as well. For all n, A_n is lower bounded by 0 and upper bounded by R_n, by application of the squeeze theorem adapted for convergence in probability, $R_n \overset{p}{\to} 0 \Rightarrow A_n \overset{p}{\to} 0$. For a short proof of how the squeeze theorem for sequences translates to convergence in probability, see Lemma A4. The assumption of strong identifiability gaurantees that $\theta \to \theta^*$ as the KL divergence goes to zero, proving the consistency of the constrained estimator. $\qquad\square$

Lemma A1. *Let $A_n(\theta), B_n(\theta) \in \mathbb{R}$ be sequences of real-valued random variables for parameter θ and $A(\theta), B(\theta)$ be random variables in $\mathbb{R}$. If both*

$$\sup_{\theta}|A_n(\theta) - A(\theta)| \overset{p}{\to} 0$$

$$\sup_{\theta}|B_n(\theta) - B(\theta)| \overset{p}{\to} 0, \tag{A8}$$

then

$$\sup_{\theta}|A_n(\theta) + B_n(\theta) - A(\theta) - B(\theta)| \overset{p}{\to} 0.$$

Proof. Define $C_n(\theta) = A_n(\theta) - A(\theta)$ and $D_n(\theta) = B_n(\theta) - B(\theta)$. By the definition of convergence in probability, the goal is to show that

$$\lim_{n\to\infty} \Pr(\sup_{\theta}|C_n(\theta) + D_n(\theta)| > \varepsilon) = 0.$$

Now construct an upper bound on the probability

$$\Pr(\sup_{\theta}|C_n(\theta) + D_n(\theta)| > \varepsilon) \leq \Pr(\sup_{\theta}|C_n(\theta)| + |D_n(\theta)| > \varepsilon)$$

$$\leq \Pr(\sup_{\theta_C}|C_n(\theta_C)| + \sup_{\theta_D}|D_n(\theta_D)| > \varepsilon)$$

$$\leq \Pr(\sup_{\theta_C}|C_n(\theta_C)| > \frac{\varepsilon}{2} \quad \text{or} \quad \sup_{\theta_D}|D_n(\theta_D)| > \frac{\varepsilon}{2}) \tag{A9}$$

$$\leq \Pr(\sup_{\theta}|C_n(\theta)| > \frac{\varepsilon}{2}) + \Pr(\sup_{\theta}|D_n(\theta)| > \frac{\varepsilon}{2})$$

$$= \Pr(\sup_{\theta}|A_n(\theta) - A(\theta)| > \frac{\varepsilon}{2}) + \Pr(\sup_{\theta}|B_n(\theta) - B(\theta)| > \frac{\varepsilon}{2}).$$

By the assumption of convergence in probability in equation (A8),

$$\lim_{n\to\infty} \Pr(\sup_{\theta}|A_n(\theta) - A(\theta)| > \frac{\varepsilon}{2}) + \Pr(\sup_{\theta}|B_n(\theta) - B(\theta)| > \frac{\varepsilon}{2}) = 0.$$

This limit shows that $\Pr(\sup_{\theta}|C_n(\theta) + D_n(\theta)| > \varepsilon)$ goes to zero via the squeeze theorem because the upper bound, equation (A9), goes to zero and the lower bound on probabilities are zero, thus showing the desired convergence in probability. $\square$

Lemma A2. *Let $M_n(\theta) \in \mathbb{R}$ be a sequence of real-valued random variables for parameter $\theta \in \Theta$ and $M(\theta), \gamma \in \mathbb{R}$ for $\theta \in \Theta$. If $\sup_{\theta \in \Theta}|M_n(\theta) - M(\theta)| \overset{p}{\to} 0$ then $\sup_{\theta \in \Theta}|(M_n(\theta) - \gamma)^2 - (M(\theta) - \gamma)^2| \overset{p}{\to} 0$.*

Proof. Define $\xi_n(\theta) = M_n(\theta) - M(\theta)$, then:

$$\sup_{\theta \in \Theta}|M_n(\theta) - M(\theta)| \overset{p}{\to} 0 \quad \Rightarrow \quad \sup_{\theta \in \Theta}|\xi_n(\theta)| \overset{p}{\to} 0. \tag{A10}$$

We have that

$$(M_n - \gamma)^2 - (M - \gamma)^2 = (M + \xi_n - \gamma)^2 - (M - \gamma)^2$$

$$= M^2 + \xi_n^2 + \gamma^2 + 2M\xi_n - 2\gamma\xi_n - 2M\gamma - M^2 - \gamma^2 + 2M\gamma \tag{A11}$$

$$= \xi_n^2 + 2\xi_n(M - \gamma).$$

We need to show that

$$\Pr(\sup_{\theta \in \Theta}|\xi_n(\theta)^2 + 2\xi_n(\theta)(M(\theta) - \gamma)| > \varepsilon) \underset{n\to\infty}{=} 0$$

This expression is the sum of two sequences with ξ_n. By Lemma A1, our desired result follows if both:

$$\sup_{\theta \in \Theta}|\xi_n(\theta)^2| \xrightarrow{p} 0$$

$$\sup_{\theta \in \Theta}|2\xi_n(\theta)(M(\theta) - \gamma)| \xrightarrow{p} 0.$$

We can bound the term with ξ_n using (A10) and the fact that $M(\theta) - \gamma \in \mathbb{R}, \forall \theta \in \Theta$. For $|\xi_n|^2$ to converge to 0, note that:

$$\Pr(\sup_{\theta \in \Theta}|\xi_n(\theta)|^2 > \varepsilon) = \Pr(\sup_{\theta \in \Theta}[|\xi_n(\theta)||\xi_n(\theta)|] > \varepsilon) = \Pr(\sup_{\theta \in \Theta}|\xi_n(\theta)| > \sqrt{\varepsilon}).$$

Again, applying (A10) gives the required result. □

Lemma A3. *Let $L_n(\theta), M_n(\theta) \in \mathbb{R}$ be sequences of real-valued random variables for $\theta \in \Theta$ and $L(\theta), M(\theta), \gamma \in \mathbb{R}$, and let $0 < \lambda < \infty$. Define*

$$\mathcal{L}_n = L_n - \lambda(M_n - \gamma)^2$$
$$\mathcal{L} = L - \lambda(M - \gamma)^2.$$

Then if $L_n \xrightarrow{p} L$ uniformly and $M_n \xrightarrow{p} M$ uniformly, then $\mathcal{L}_n \xrightarrow{p} \mathcal{L}$ uniformly,

Proof. By Lemma A2, we have that

$$\sup_{\theta}|M_n(\theta) - M(\theta)| \xrightarrow{p} 0 \implies \sup_{\theta \in \Theta}|(M_n(\theta) - \gamma)^2 - (M(\theta) - \gamma)^2| \xrightarrow{p} 0.$$

By uniform convergence in probability, we have that also

$$\sup_{\theta}|L_n(\theta) - L(\theta)| \xrightarrow{p} 0.$$

This gives us

$$\sup_{\theta}|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| = \sup_{\theta}|L_n(\theta) - \lambda(M_n(\theta) - \gamma)^2 - L + \lambda(M(\theta) - \gamma)^2|$$

$$\overset{i}{\leq} \sup_{\theta}|L_n(\theta) - L(\theta)| + \lambda\sup_{\theta}|(M(\theta) - \gamma)^2 - (M_n(\theta) - \gamma)^2| \xrightarrow{p} 0.$$

(i) follows from the triangle inequality and the final convergence follows from $0 < \lambda < \infty$ and the uniform convergence of the individual terms. □

Lemma A4. *If $X_n \leq Y_n \leq Z_n$ for every n and $X_n \xrightarrow{p} L$ and $Z_n \xrightarrow{p} L$, then $Y_n \xrightarrow{p} L$*

Proof. We want to show that for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty}\Pr(|Y_n - L| > \varepsilon) = 0.$$

Note that

$$X_n \leq Y_n \leq Z_n$$

implies

$$|Y_n - L| \leq |X_n - L| + |Z_n - L|,$$

so

$$\Pr(|Y_n - L| > \varepsilon) \leq \Pr(|X_n - L| + |Z_n - L| > \varepsilon).$$

The right-hand side can be made arbitrarily small by convergence of X_n and Z_n, proving what we want to show. □

A.3 The offset model is not a consistent estimator in the presence of unobserved confounding

We now prove that optimization over the offset model family in equation (7) and a known conditional odds ratio β_t leads to an asymptotically biased estimator of $\Pr(Y_t|x)$ in the presence of unobserved confounding in a simple example introduced in the main text (4.1.1). In this example, $\Pr(Y_t|x) = \Pr(Y_t)$, or equivalently, there is no covariate X. This also implies that here CATE = ATE, so we will use the ATE instead. Given the offset model family from equation (7), a natural parameterization of $g(t, x) = g(t)$ in the context of this example is $g(t; b_0) = \sigma(b_0 + \beta_t t)$, $b_0 \in \mathbb{R}$. Again, we are assuming that β_t is given *a priori* and is not estimated. We first derive an expression for the expected log-likelihood as a function of b_0 under the observational distribution in this example. Then we show that the ground truth solution β_0 is not a stationary point, proving our claim. Writing

$$p_u = p(u = 1)$$
$$p_{t'u'} = p(t = t', u = u') = p(t = t'|u = u')p(u = u')$$
$$\pi_{t'u'} = p(y = 1|t = t', u = u').$$

Then the data-generating mechanism is:

$$t, u \sim \mathcal{B}(p_{t'u'}), y \sim \mathcal{B}(\pi_{tu}).$$

The ground truth solutions β_0 and β_t are:

$$\Pr(Y_0 = 1) = (1 - p_u)\pi_{00} + p_u\pi_{01} = \sigma(\beta_0)$$
$$\Pr(Y_1 = 1) = (1 - p_u)\pi_{10} + p_u\pi_{11} = \sigma(\beta_0 + \beta_t). \tag{A12}$$

The Bernoulli log-likelihood is

$$l(y|t, b_0) = y\log\sigma(b_0 + \beta_t t) + (1 - y)\log(1 - \sigma(b_0 + \beta_t t))$$

In offset models, β_t is assumed given *a priori* and b_0 is the only parameter, resulting in the following expression for $L(b_0)$:

$$L(b_0) = p_{00}[\pi_{00}\log\sigma(b_0) + (1 - \pi_{00})\log(1 - \sigma(b_0))] + p_{01}[\pi_{01}\log\sigma(b_0) + (1 - \pi_{01})\log(1 - \sigma(b_0))]$$
$$+ p_{10}[\pi_{10}\log\sigma(b_0 + \beta_t) + (1 - \pi_{10})\log(1 - \sigma(b_0 + \beta_t))]$$
$$+ p_{11}[\pi_{11}\log\sigma(b_0 + \beta_t) + (1 - \pi_{11})\log(1 - \sigma(b_0 + \beta_t))].$$

Taking the derivative with respect to b_0, noting that $(\log\sigma(x))' = 1 - \sigma(x)$, we obtain:

$$\frac{\partial L}{\partial b_0} = p_{00}[\pi_{00}(1 - \sigma(b_0)) - (1 - \pi_{00})\sigma(b_0)] + p_{01}[\pi_{01}(1 - \sigma(b_0)) - (1 - \pi_{01})\sigma(b_0)]$$
$$+ p_{10}[\pi_{10}(1 - \sigma(b_0 + \beta_t)) - (1 - \pi_{10})\sigma(b_0 + \beta_t)] \tag{A13}$$
$$+ p_{11}[\pi_{11}(1 - \sigma(b_0 + \beta_t)) - (1 - \pi_{11})\sigma(b_0 + \beta_t)].$$

We now plug in the ground truth solutions for β_0, β_t (equation (A12)).

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0) = p_{00}[\pi_{00}(1 - p_u\pi_{01} - (1 - p_u)\pi_{00}) - (1 - \pi_{00})(p_u\pi_{01} + (1 - p_u)\pi_{00})]$$
$$+ p_{01}[\pi_{01}(1 - p_u\pi_{01} - (1 - p_u)\pi_{00}) - (1 - \pi_{01})(p_u\pi_{01} + (1 - p_u)\pi_{00})]$$
$$+ p_{10}[\pi_{10}(1 - p_u\pi_{11} - (1 - p_u)\pi_{10}) - (1 - \pi_{10})(p_u\pi_{11} + (1 - p_u)\pi_{10})]$$
$$+ p_{11}[\pi_{11}(1 - p_u\pi_{11} - (1 - p_u)\pi_{10}) - (1 - \pi_{11})(p_u\pi_{11} + (1 - p_u)\pi_{10})].$$

Removing terms that cancel out in each line results in

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0) = p_{00}[p_u(\pi_{00} - \pi_{01})] + p_{01}[(1 - p_u)(\pi_{01} - \pi_{00})]$$
$$+ p_{10}[p_u(\pi_{10} - \pi_{11})] + p_{11}[(1 - p_u)(\pi_{11} - \pi_{10})].$$

Substituting back $p_{t'u'} = p(t = t'|u = u')p(u = u')$:

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0) = p(t = 0|u = 0)(1 - p_u)[p_u(\pi_{00} - \pi_{01})]$$
$$+ p(t = 0|u = 1)p_u[(1 - p_u)(\pi_{01} - \pi_{00})]$$
$$+ p(t = 1|u = 0)(1 - p_u)[p_u(\pi_{10} - \pi_{11})]$$
$$+ p(t = 1|u = 1)p_u[(1 - p_u)(\pi_{11} - \pi_{10})].$$

Factoring out $p_u(1 - p_u)$ and re-arranging, we arrive at our result:

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0) = p_u(1 - p_u)[(\pi_{01} - \pi_{00})(p(t = 0|u = 1) - p(t = 0|u = 0))$$
$$+ (\pi_{11} - \pi_{10})(p(t = 1|u = 1) - p(t = 1|u = 0))].$$

If there is no confounding ($\pi_{t0} = \pi_{t1}$ or $p(t|u = 0) = p(t|u = 1)$), this expression is zero, but in general, it is not which means that the ground truth solution β_0 is not an optimum of the expected log-likelihood in the observational data distribution. This proves our claim that the offset model does not recover $\Pr(Y_t)$ in the presence of confounding. □

Of note, the fact that the offset model does not estimate $\Pr(Y_t)$ does not automatically imply that the ATE is not correctly estimated as there may be another $b_0' \neq \beta_0$ such that $\widehat{ATE}(b_0 = b_0') = ATE(b_0 = \beta_0)$. To investigate this, assume that for some $\beta_0 = a$ and $\beta_t = c$, we have that:

$$\delta := \sigma(a + c) - \sigma(a) = \frac{e^{a+c}}{1 + e^{a+c}} - \frac{e^a}{1 + e^a}.$$

Again, by treating β_t as fixed, we will now prove that this equation has at most two solutions for $b_0 = a$ by noting that:

$$\frac{e^{a+c}}{1 + e^{a+c}} - \frac{e^a}{1 + e^a} = \frac{e^{a+c}(1 + e^a) - (1 + e^{a+c})e^a}{(1 + e^{a+c})(1 + e^a)} = \frac{e^a(e^c - 1)}{(1 + e^{a+c})(1 + e^a)}$$

By introducing $y := e^a$ and cross-multiplying, we obtain:

$$\delta = \frac{y(e^c - 1)}{(1 + e^c y)(1 + y)} \quad \Leftrightarrow \quad \delta(1 + e^c y)(1 + y) = y(e^c - 1)$$
$$= \delta + \delta(1 + e^c)y + \delta e^c y^2 = y(e^c - 1)$$
$$\Leftrightarrow \delta e^c y^2 + (\delta(1 + e^c) - e^c + 1)y + \delta = 0.$$

Depending on the values of δ and c, this quadratic equation in y has 0, 1 or 2 real-valued solutions, yielding 0, 1 or 2 real-valued solutions for $a = \log y = b_0$. This implies that there exists utmost one alternative solution $b_0' \neq \beta_0$ such that $\widehat{ATE}(b_0 = b_0') = ATE(b_0 = \beta_0)$.

In fact, we can explicitly compute this alternative solution by exploiting the symmetry of the sigmoid function: $\sigma(x) = 1 - \sigma(-x)$. Whenever it is true that:

$$\sigma(\beta_0 + \beta_t) - \sigma(\beta_0) = \delta.$$

It must simultaneously be true that, writing $\beta_0' := -(\beta_0 + \beta_t)$:

$$\sigma(\beta_0' + \beta_t) - \sigma(\beta_0') = \sigma(-(\beta_0 + \beta_t) + \beta_t) - \sigma(-(\beta_0 + \beta_t))$$
$$= \sigma(-\beta_0) - \sigma(-(\beta_0 + \beta_t))$$
$$= (1 - \sigma(\beta_0)) - (1 - \sigma(\beta_0 + \beta_t))$$
$$= \sigma(\beta_0 + \beta_t) - \sigma(\beta_0) = \delta.$$

This means that except in the trivial case when $\beta_0 = \beta_t = 0$ there *always* exists a second solution β_0' that has the same ATE δ but a different $\Pr(Y_t)$. We can check whether this coincidentally coincides with the maximum likelihood solution for b_0 in the offset model on the observational data by plugging in $\beta_0' := -(\beta_0 + \beta_t)$ in the expression of the gradient of the likelihood (equation (A13)). Again we remove terms that cancel out and substitute back $p_{t'u'} = p(t = t'|u = u')p(u = u')$ to arrive at:

$$\frac{\partial L}{\partial b_0}(b_0 = \beta_0') = p_u(1 - p_u)\big(p(t = 0|u = 1) - p(t = 0|u = 0)\big)\big((\pi_{10} - \pi_{11}) + (\pi_{01} - \pi_{00})\big) \tag{A14}$$

$$+p_u\big((\pi_{11} - \pi_{10}) + (\pi_{01} - \pi_{00})\big) \tag{A15}$$

$$+2\pi_{10} + \pi_{11} - 1. \tag{A16}$$

By analyzing this expression line-by-line, we see that the first two lines are non-zero in general when there is confounding such that $p(t = 0|u = 1) \neq p(t = 0|u = 0)$ and $\pi_{t1} \neq \pi_{t0}$. The last line is also non-zero in general as π_{tu} are free parameters. This shows that the offset model is also an asymptotically biased estimator of the ATE.

A.4 Analytical solution of offset model with binary covariate

Assume a binary treatment t, covariate x and outcome y. Write:

$$\beta_t^0 = \sigma^{-1}(\Pr(Y_1 = 1|x = 0)) - \sigma^{-1}(\Pr(Y_0 = 1|x = 0)), \tag{A17}$$

$$\beta_t^1 = \sigma^{-1}(\Pr(Y_1 = 1|x = 1)) - \sigma^{-1}(\Pr(Y_0 = 1|x = 1)). \tag{A18}$$

The offset assumption states that $\beta_t^0 = \beta_t^1 = \beta_t$.

Maximizing the likelihood over a model class is equivalent to minimizing the Kullback–Leibler divergence between the actual distribution and the model distribution. We find an analytical solution for the offset models by writing the estimation objective as minimizing the Kullback–Leibler divergence between the model distribution and the observed data distribution with the additional constraints given in A17. Writing:
- p_{tx} for estimated probabilities of $\Pr(Y = 1|t, x)$, such that $p(y|t, x) = (1 - y)(1 - p_{tx}) + yp_{tx}$
- q_{tx} for actual (observational) probabilities of $\Pr(Y = 1|t, x)$, such that $q(y|t, x) = (1 - y)(1 - q_{tx}) + yq_{tx}$
- $q(t, x) = \Pr(T = t, X = x)$, the observational joint probability of T, X.

The criterion is:

$$L = \mathbb{E}_{t,x}\mathbb{E}_{y|t,x}\log\frac{q(y|t, x)}{p(y|t, x)} + \lambda_0(\sigma^{-1}(p_{10}) - \sigma^{-1}(p_{00}) - \beta_t) + \lambda_1(\sigma^{-1}(p_{11}) - \sigma^{-1}(p_{01}) - \beta_t). \tag{A19}$$

The partial derivatives for all parameters $(p_{00}, p_{01}, p_{10}, p_{11}, \lambda_0, \lambda_1)$ are as follows:

$$\frac{\partial L}{\partial p_{00}} = \mathbb{E}_{t,x}\mathbb{E}_{y|t,x}\frac{\partial}{\partial p_{00}}[(\log q(y|t, x) - \log p(y|t, x)) + \lambda_0(\sigma^{-1}(p_{10}) - \sigma^{-1}(p_{00}) - \beta_t) \\ + \lambda_1(\sigma^{-1}(p_{11}) - \sigma^{-1}(p_{01}) - \beta_t)], \tag{A20}$$

$$= \mathbb{E}_{t,x}\mathbb{E}_{y|t,x}\frac{\partial}{\partial p_{00}}[(\log q(y|t, x) - \log(yp_{tx} + (1 - y)(1 - p_{tx}))] - \lambda_0\frac{\partial}{\partial p_{00}}\sigma^{-1}(p_{00}), \tag{A21}$$

$$= q(0, 0)\left[(1 - q_{00})\frac{1}{1 - p_{00}} - q_{00}\frac{1}{p_{00}}\right] - \lambda_0\left(\frac{1}{p_{00}} + \frac{1}{1 - p_{00}}\right), \tag{A22}$$

$$= q(0, 0)\left[\frac{(1 - q_{00})p_{00} - q_{00}(1 - p_{00})}{p_{00}(1 - p_{00})}\right] - \lambda_0\left(\frac{(1 - p_{00}) + p_{00}}{p_{00}(1 - p_{00})}\right), \tag{A23}$$

$$= q(0, 0)\left[\frac{p_{00} - q_{00}}{p_{00}(1 - p_{00})}\right] - \lambda_0\left(\frac{1}{p_{00}(1 - p_{00})}\right), \tag{A24}$$

$$= \frac{q(0,0)(p_{00} - q_{00}) - \lambda_0}{p_{00}(1 - p_{00})}, \tag{A25}$$

$$\frac{\partial L}{\partial p_{01}} = \frac{q(0,1)(p_{01} - q_{01}) - \lambda_1}{p_{01}(1 - p_{01})}, \tag{A26}$$

$$\frac{\partial L}{\partial p_{10}} = \frac{q(1,0)(p_{10} - q_{10}) + \lambda_0}{p_{10}(1 - p_{10})}, \tag{A27}$$

$$\frac{\partial L}{\partial p_{11}} = \frac{q(1,1)(p_{11} - q_{11}) + \lambda_1}{p_{11}(1 - p_{11})}, \tag{A28}$$

$$\frac{\partial L}{\partial \lambda_0} = \sigma^{-1}(p_{10}) - \sigma^{-1}(p_{00}) - \beta_t, \tag{A29}$$

$$\frac{\partial L}{\partial \lambda_1} = \sigma^{-1}(p_{11}) - \sigma^{-1}(p_{01}) - \beta_t. \tag{A30}$$

By setting the gradient to zero to find an extremum of A19 and requiring $0 < p_{tx} < 1$, we obtain:
$0=$

(a) $q(0,0)(p_{00} - q_{00}) - \lambda_0$,
(b) $q(0,1)(p_{01} - q_{01}) - \lambda_1$,
(c) $q(1,0)(p_{10} - q_{10}) + \lambda_0$,
(d) $q(1,1)(p_{11} - q_{11}) + \lambda_1$,
(e) $\sigma^{-1}(p_{10}) - \sigma^{-1}(p_{00}) - \beta_t$,
(f) $\sigma^{-1}(p_{11}) - \sigma^{-1}(p_{01}) - \beta_t$.

Combining (a) and (c):

$$a + c = q(0,0)(p_{00} - q_{00}) + q(1,0)(p_{10} - q_{10}) - \lambda_0 + \lambda_0 = 0 \tag{A31}$$

$$0 = q(0,0)(p_{00} - q_{00}) + q(1,0)(p_{10} - q_{10}). \tag{A32}$$

By inserting (e): $p_{10} = \sigma(\sigma^{-1}(p_{00}) + \beta_t)$ we obtain

$$0 = q(0,0)(p_{00} - q_{00}) + q(1,0)(\sigma(\sigma^{-1}(p_{00}) + \beta_t) - q_{10}) \tag{A33}$$

By using $\sigma(\sigma^{-1}(x) + a) = \frac{e^a x}{(e^a - 1)x + 1}$, we obtain:

$$0 = q(0,0)(p_{00} - q_{00}) + q(1,0)\left[\frac{e^{\beta_t} p_{00}}{(e^{\beta_t} - 1)p_{00} + 1} - q_{10}\right] \tag{A34}$$

By multiplying both sides with $(e^{\beta_t} - 1)p_{00} + 1$, we end up with a quadratic equation in p_{00}. Using some intermediate variables to shorten the notation and applying the same steps to obtain p_{01}, the solution is:

$$\tau = e^{\beta_t}$$
$$r_x = q(1,x)/q(0,x)$$
$$c_x = -(q(y|0,x) + r_x q(y|1,x))$$
$$b_x = 1 + r_x \tau + c_x(\tau - 1)$$
$$p_{0x} = \frac{-b_x \pm \sqrt{b_x^2 + 4(\tau - 1)c_x}}{2(\tau - 1)}.$$

The $\pm$ solution is taken to equal sign(β_t). Without loss of generality, we assume $\beta_t > 0$. The corresponding solutions p_{1x} are given by $p_{1x} = \sigma(\sigma^{-1}(p_{0x}) + \beta_t)$.
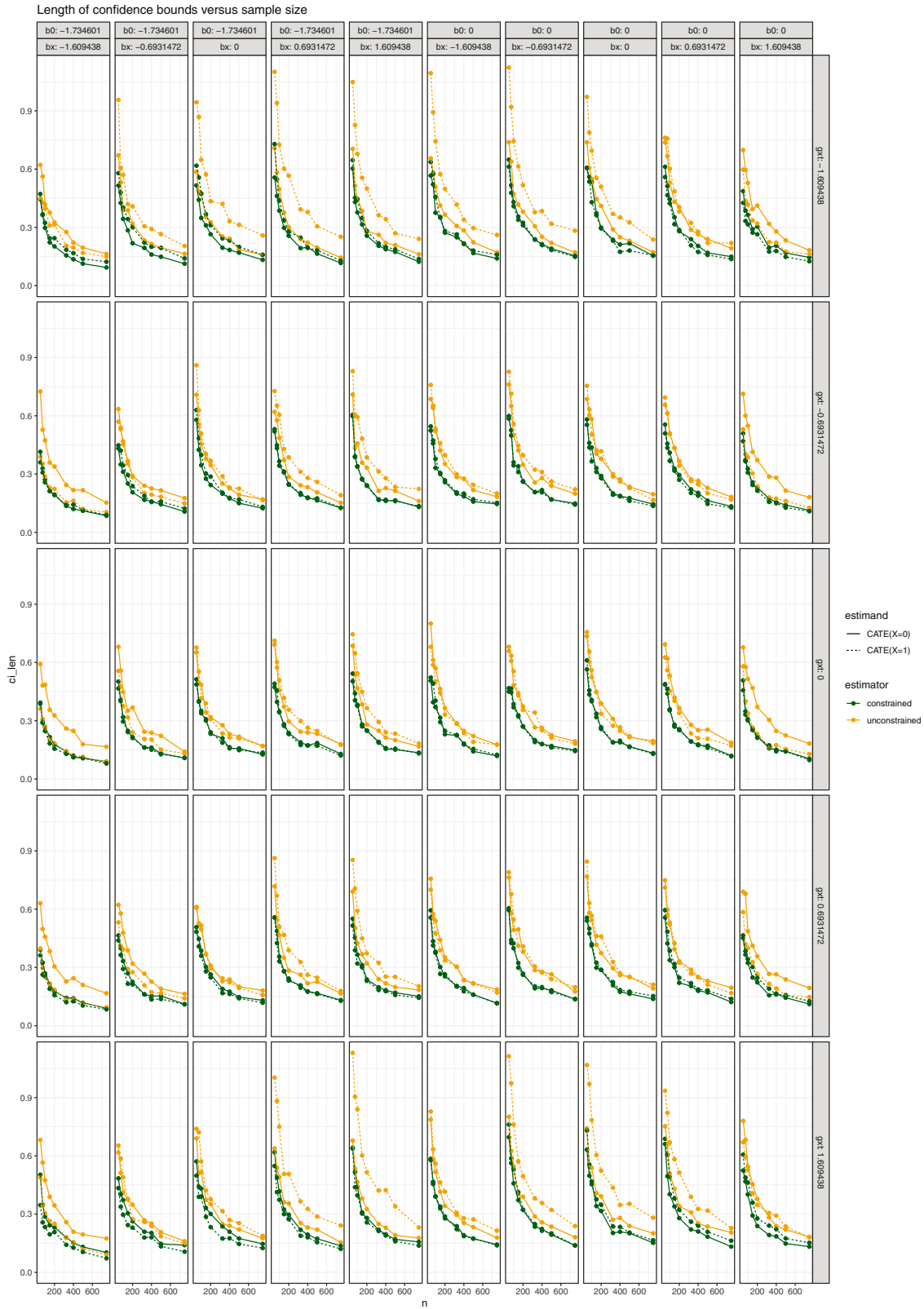
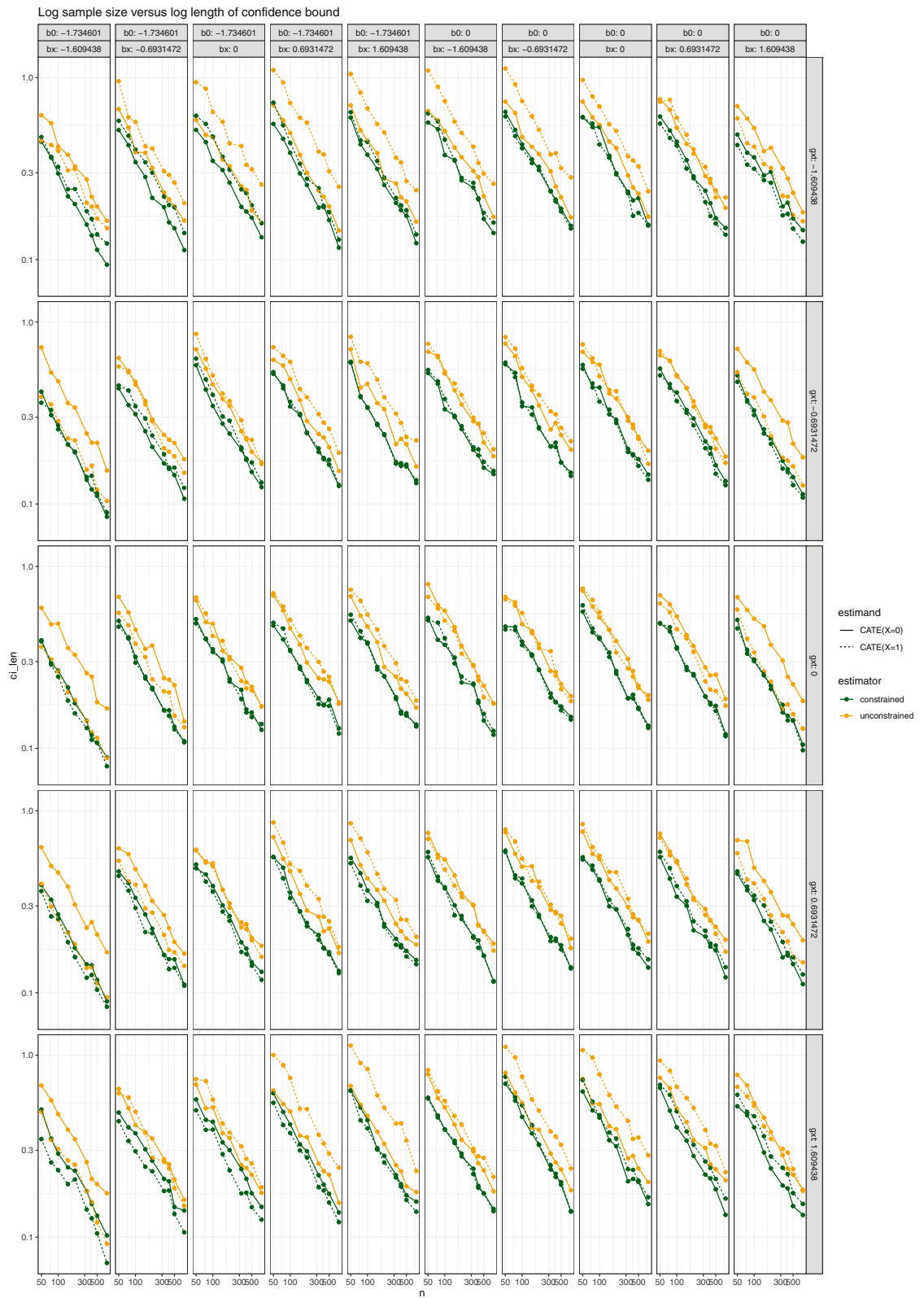**Figure A2:** Results of efficiency experiment (Section 5.1).

Figure A3: Results of efficiency experiment on a log-scale (Section 5.1).