Inference and estimation with unidentifiable latent treatment effects

Christopher Tosh
Computational Oncology
Memorial Sloan Kettering Cancer Center

Boyuan Zhang
Department of Statistics
Stanford University

Wesley Tansey
Computational Oncology
Memorial Sloan Kettering Cancer Center

Abstract

In experimental and clinical studies, samples that responded to treatment are often analyzed in order to refine scientists' hypotheses and find potential causal drivers of response. Natural variation in outcomes makes teasing apart responders from non-responders a statistical inference problem. We introduce the causal two-groups (C2G) model, a causal extension of the classical two-groups model for inferring treatment effects. The classical two-groups model posits that treated samples may or may not experience an effect, according to some prior probability. The C2G model extends this to consider the case where treatment, latent effect status, and effect size are all confounded. We propose two empirical Bayes procedures for the causal two-groups model, one under semi-parametric conditions and another under fully nonparametric conditions. The semi-parametric model assumes additive treatment effects and is identifiable from observed data. The nonparametric model is unidentifiable, but we show it can still be used to test for response in each treated sample. We show empirically and theoretically that both methods for selecting responders control the false discovery rate at the target level with near-optimal power. We also propose two novel estimands of interest and provide a strategy for deriving estimand intervals in the nonparametric model. We apply the C2G model to a cancer immunotherapy dataset, where the nonparametric model recovers clinically validated predictive biomarkers of both positive and negative outcomes.[1]

1 Introduction

In a dataset with treatment and control groups, what effects can be inferred if only an unknown subset of the treated subjects receive an effect? Can we identify treated individuals that saw an effect from the treatment? Can we estimate the magnitudes of effects? What kinds of statistical guarantees can we provide?

Such questions commonly arise in medicine, where only individuals with certain characteristics experience the benefits of an intervention. For example, a subpopulation of Hispanic patients has a germline mutation that prevents an approved CAR-T cell therapy from binding its target [46]. Lung tumors often acquire a somatic mutation that prevents certain small molecule inhibitors from binding [7]. Personalized cancer mRNA vaccines are only effective if the neoantigens targeted provoke an immune response [41]. In all of these cases, biomarkers of response were unknown at the start of the study. The key step in refining the treatment criteria was to determine which patients

[1] Code is available at `https://github.com/tansey-lab/causal2groups`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

likely experienced an effect on their outcomes. Since patients experience variability in their outcomes, determining which patients were affected and by how much is a statistical inference problem.

Latent treatment responses also arise in the intention-to-treat setting, where a treatment is assigned but compliance with treatment is unmeasured. This setting occurs whenever a subject has to elect to take the treatment, such as observational studies of outpatient health record analyses or voucher efficacy analyses in social welfare programs [17]. Traditional approaches to intention-to-treat analysis use the treatment assignment variable as the causal variable, despite being a biased estimator of the true treatment effect. Estimating effects among compliers has previously only been feasible when either compliance is observed [12, 14] or instrumental/proxy variables are available [36, 49, 4]. This paper explores the limits and capabilities of causal inference with fully unobserved compliance and no available proxies or instruments. In doing so, we provide new tools to make causal conclusions in the face of latent treatment compliance even under certain forms of latent confounding. Hereon, we will use the latent treatment response terminology, but all results also apply to the equivalent intention-to-treat/noncompliance interpretation.

Here, we develop the causal two-groups (C2G) model, an empirical Bayes framework extending the classical two-groups model [10] to the latent treatment response setting (Figure 1e). The C2G model assumes a binary treatment T, binary latent response H, continuous outcome Y, and observed confounders X, with extensions to unmeasured confounders U (Section 5). When H is fully unobserved and outcome distributions are nonparametric, the model is unidentifiable in general (Proposition 1). Previous works circumvent this by assuming H is observable or by imposing parametric assumptions. We show two perhaps surprising results: (1) a flexible semi-parametric C2G model is fully identifiable and can be fit via EM, and (2) despite unidentifiability, the nonparametric model still permits non-trivial *interval* estimates of causal effects. These results allow us to select individuals who responded to treatment with false discovery rate (FDR) guarantees and near-optimal power, to construct informative intervals for two new estimands previously considered unidentifiable, and to guarantee robustness to certain forms of latent confounding. We demonstrate practical applicability in a cancer immunotherapy case study.

1.1 Related work

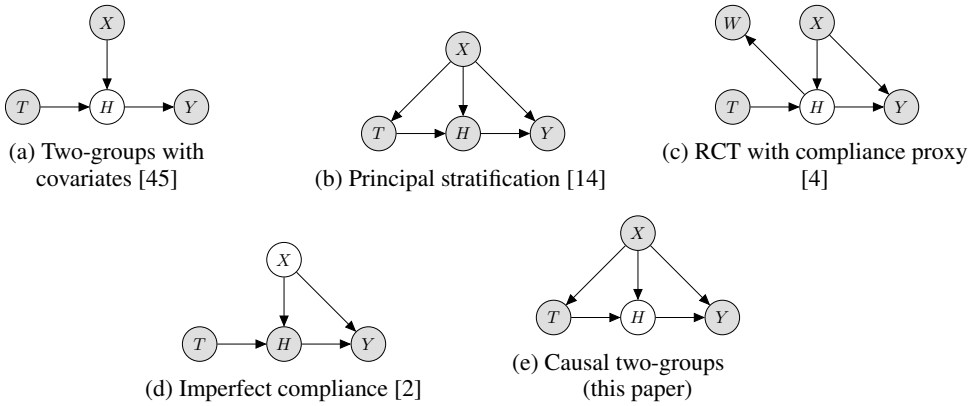
The causal two-groups model draws on and connects several lines of research. Here we provide a short discussion on the closest related work. An extended discussion is presented in Appendix A.

Multiple testing with covariates. Several approaches to multiple testing incorporate covariates as side information. These methods generally work under the two-groups model [9, 10, 11] where a test statistic is drawn either from the null or alternative with some prior probability. However, both frequentist [20, 56, 27, 5, 28] and Bayesian [45, 37, 50, 51] approaches assume that the covariates only impact the likelihood of treatment having an effect. This is often a reasonable assumption in laboratory experiments, where biological replicates can be tested repeatedly under different experimental conditions. However, it is less likely to hold in observational data where covariates are often confounders that affect the probability of treatment assignment, treatment compliance, and the outcomes themselves. Figure 1a shows the graphical model for existing two-groups models.

Principal stratification and noncompliance. Principal stratification provides a framework for reasoning about causal effects when post-treatment intermediate variables are present [14]. In the compliance setting, prior work has required either that compliance is directly observed (Figure 1b), that proxy or instrumental variables are available (Figure 1c), or that strong parametric assumptions hold [1, 4]. The C2G model addresses the more challenging regime where the latent response variable is fully unobserved and no instruments or proxies are available. Unlike classical principal stratification, we show that meaningful inference, including FDR-controlled selection of responders and informative estimand intervals, is possible even when the model is not fully identifiable.

2 The causal two-groups model

Let $(x_i, y_i, t_i)_{i=1}^n$ be an observational dataset of n samples, where $t_i \in \{0, 1\}$ indicates whether individual i was treated $(t_i = 1)$ or not $(t_i = 0)$, $y_i \in \mathbb{R}$ is the individual outcome, and $x_i \in \mathbb{R}^d$ are potential confounders. For instance, $t_i = 1$ indicates that the patient was prescribed a certain drug, y_i



(a) Two-groups with covariates [45]

(b) Principal stratification [14]

(c) RCT with compliance proxy [4]

(d) Imperfect compliance [2]

(e) Causal two-groups (this paper)

Figure 1: (a) The FDR regression model of Scott et al. [45]. (b) An observational dataset with observed noncompliance [14]. (c) A randomized controlled trial with a compliance proxy [4]. (d) A randomized controlled trial with imperfect compliance and no proxy [2]. (e) The general causal two-groups model. The treatment T, (unobserved) response indicator H, and the outcome Y are all confounded by X.

is the patient's survival time in days, and x_i is demographic information like age, race, and sex. We drop the subscripts from here on when it is clear we are referring to a single observation.

Our primary focus will be on selecting responders, but we will also consider estimation of effect sizes. We adapt the two-groups model from the multiple testing literature [9, 10]. For each treated individual ($t = 1$), we suppose they are either a *non-responder* or a *responder*, with outcome distributions $f_0(y \mid x)$ or $f_1(y \mid x)$, respectively. All untreated ($t = 0$) patients are assumed to have the same outcome distribution as non-responders. Each treated individual is a responder with probability $\pi(x)$. In probabilistic modeling terms,

$$\begin{aligned} y \mid x, h &\sim f_h(y \mid x), & h \mid x, t{=}0 &= 0, \\ h \mid x, t{=}1 &\sim \text{Bern}(\pi(x)), & t \mid x &\sim \text{Bern}(\phi(x)), \end{aligned} \tag{1}$$

where $h \in \{0, 1\}$ indicates whether the individual responded to treatment. Figure 1e shows the causal graphical model corresponding to Equation (1). The model assumes X is a confounder of T, H, and Y. Scenarios with no treatment bias (e.g., randomized controlled trials) are covered as well, as Equation (1) is strictly a generalization. The model also assumes that treatment response is random, allowing for exogenous variables (e.g., unmeasured biomarkers) to induce uncertainty in response.

2.1 When is the causal two-groups model identifiable?

Our first result is that, with no further assumptions, the causal two-groups model is not identifiable, even without overlap violations.

Proposition 1. *The model in Equation* (1) *is unidentifiable without further assumptions, even when* $0 < \pi(x) < 1$ *for all* x*. Moreover, it is unidentifiable even when* x *is a constant and the distributions* f_0, f_1 *are restricted to being differentiable log-concave probability distributions.*

For space and clarity, all proofs are presented in the appendix. The key idea is to rewrite the treatment outcome distribution $f_t(y \mid x) := p(y \mid x, t = 1)$ as a mixture model,

$$f_t(y \mid x) = (1 - \pi(x))f_0(y \mid x) + \pi(x)f_1(y \mid x). \tag{2}$$

Then one can show that even when $f_0(y \mid x)$ is fixed, there is a range of π's and f_1's that give rise to exactly the same f_t.

The story changes if parametric forms of the conditional likelihoods f_0, f_1 are known. Indeed, the form of Equation (2) implies that Equation (1) is identifiable at those x such that $f_0(\cdot \mid x), f_1(\cdot \mid x)$ come from a family for which finite mixtures are identifiable [52, 57]. As a concrete example, we give a direct proof that Equation (1) is identifiable for the class of conditional normal likelihoods.

Proposition 2. *Suppose that* f_h *is of the form* $f_h(y \mid x) = \mathcal{N}(y \mid \mu_h(x), \sigma_h^2(x))$ *for* $h = 0, 1$*. Then Equation* (1) *is identifiable for all* x *such that* $f_0(\cdot \mid x) \neq f_1(\cdot \mid x)$*.*

In practice, we generally do not know the parametric forms of the outcome distributions a priori. Thus, we would ideally like to avoid parametric assumptions as much as possible. As we show in Sections 3 and 4, there is some hope here: nonparametric estimation of the true model may be impossible, but semi-parametric estimation and nonparametric testing are viable.

2.2 Testing in the causal two-groups model

We formulate responder selection as a multiple hypothesis testing problem. That is, for each treated sample, we test $H_0 \colon h_i = 0$ (i.e., the outcome was drawn from the non-responder distribution), aiming to maximize power while controlling the false discovery rate, $\text{FDR} = \mathbb{E}\left[\frac{1}{|\hat{V}|}\sum_{i\in\hat{V}} \mathbb{1}[h_i = 0]\right]$, at or below a target level α.

Following the empirical Bayes approach common to two-groups models [9, 10, 11, 45, 50, 51], our procedure has four steps. First, we fit $\hat{f}_0(y \mid x)$ on the untreated population, which by Equation (1) also models non-responder outcomes. Second, we estimate $\hat{f}_1(y \mid x)$ and $\hat{\pi}(x)$ from the treated population. Third, we compute conservative (upper-bound) posterior non-response probabilities

$$\hat{w}_i = \frac{(1-\hat{\pi}(x_i))\hat{f}_0(y_i \mid x_i)}{(1-\hat{\pi}(x_i))\hat{f}_0(y_i \mid x_i) + \hat{\pi}(x_i)\hat{f}_1(y_i \mid x_i)}\,.$$

Finally, we select discoveries by solving

$$\underset{V}{\text{maximize}}\ |V| \quad \text{subject to} \quad \frac{1}{|V|}\sum_{i\in V} \hat{w}_i \le \alpha\,. \tag{3}$$

The modeling choices for f_0, f_1, and π depend on the assumptions one is willing to make. Section 3 presents a semi-parametric model assuming additive noise with nonparametric mean functions and noise distribution. Section 4 considers fully nonparametric outcome distributions.

2.3 Estimands of interest in the causal two-groups model

Latent (non-)responders make interpreting causal effects more nuanced. Here, we introduce two estimands of potential interest to the analyst.

The *conditional average response effect (CARE)*, analogous to the conditional average treatment effect (CATE), captures the treatment impact among responders:

$$\text{CARE}(x) := \mathbb{E}[Y \mid T = 1, H = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x]\,.$$

For positive effects, the CATE will generally be a biased downward estimate of the CARE. Removing conditioning on X yields the average response effect (ARE). The *expected responder population fraction (ERPF)* captures how many treated individuals experienced an effect:

$$\text{ERPF} := P(H = 1 \mid T = 1).$$

Together, the CARE and ERPF distinguish between treatments with a strong effect for few patients versus a small but consistent effect across the population. Our ability to estimate these quantities depends on whether the modeling assumptions render them identifiable or allow for interval estimation. When only an interval is identifiable (Section 4), we use the end closer to zero to provide conservative estimates. Section 7 demonstrates estimation of both estimands on a cancer immunotherapy dataset.

2.4 Potential outcomes interpretations of the causal two-groups model

Although the selection procedures and interval estimates in this paper depend only on the core assumption of Equation (1), it is possible to give a counterfactual interpretation of the quantities arising in the C2G model under additional assumptions. To do so, recall the potential outcomes framework [43] in which each individual has a potential outcome $Y(t)$ for each possible treatment $t \in \{0, 1\}$, and we observe the outcome $Y = Y(T)$. Analogously, each individual also has potential response variables $H(t) \in \{0, 1\}$ for each $t \in \{0, 1\}$, and the realized response is $H = H(T)$. $H(1)$ indicates whether the individual would respond under treatment, and $H(0)$ indicates whether they would respond without it. Mirroring the constraint of Equation (1) that untreated individuals cannot respond, we will assume here that $H(0) = 0$, which implies that $H = T \cdot H(1)$.

Given this language, we can formalize our causal assumptions.

(A1) **Unconfounded assignment.** $(Y(0), Y(1), H(1)) \perp\!\!\!\perp T \mid X$, i.e., none of the potential variables depend on the treatment variable conditioned on X.

(A2) **Exclusion restriction.** $H(1)=0 \Rightarrow Y(1)=Y(0)$, i.e., treatment only affects an individual that responds.

(A3) **Principal ignorability.** $Y(0) \perp\!\!\!\perp H(1) \mid X$, i.e., given X, responders and non-responders have the same distribution of untreated outcomes.

Under these assumptions, we can make the following statement.

Theorem 3. *Under assumptions (A1) and (A3),*

$$\text{CARE}(x) = \mathbb{E}[Y(1) - Y(0) \mid H(1) = 1, X = x].$$

Under assumption (A2),

$$\text{ERPF} = P(H(1) = 1 \mid T = 1) \geq P(Y(0) \neq Y(1) \mid T = 1).$$

In the principal stratification literature [14], individuals are grouped by their potential responses ($H(0)$ and $H(1)$, here), which do not depend on the treatment they actually receive. In the causal two-groups model, we assume $H(0) = 0$, and so there are only two strata consisting of responders ($H(1) = 1$) and non-responders ($H(1) = 0$). Thus, the first part of Theorem 3 establishes that CARE is the principal causal effect in the responder stratum. The second part of Theorem 3 shows that ERPF is the size of the responder stratum conditioned on treatment and that it is lower bounded by the probability of treated individuals seeing an effect from treatment.

The potential outcomes framing also allows us to understand the causal two-groups setting through the lens of compliance. Here, T is treatment assignment, and H is the unobserved treatment compliance. The assumption of $H(0) = 0$ is equivalent to one-sided noncompliance, i.e., individuals not assigned treatment cannot get access to treatment. Thus, the two principal strata here are compliers ($H(1) = 1$) and never-takers ($H(1) = 0$). Then the consequences of Theorem 3 are that CARE is the average effect of treatment assignment among compliers and that ERPF is the fraction of compliers in the set assigned treatment.

3 A semi-parametric additive errors model

In the additive errors (AE) regime, we model y as a deterministic function plus i.i.d. noise,

$$(y \mid x, h = 0) = \mu_0(x) + \epsilon, \quad (y \mid x, h = 1) = \mu_1(x) + \epsilon, \quad \epsilon \sim g(\epsilon). \tag{4}$$

Equation (4) is common in many causal inference setups, particularly when estimating the conditional average treatment effect [cf. 16, 53]. This setting can be thought of as a conditional semi-parametric model: conditioned on covariates x, the model consists of the parameters $\mu_0(x), \mu_1(x)$ and the infinite-dimensional model for the noise ϵ. The AE model is strictly more flexible than a location-only parametric model. Despite this added flexibility, the model remains identifiable, meaning we can directly estimate the underlying parameters and distribution.

Theorem 4. *The model in Equation* (4) *is identifiable at every x such that $\mu_1(x) \neq \mu_0(x)$ and $0 < \pi(x) < 1$.*

The upshot of Theorem 4 is that we can fit a semi-parametric model to Equation (4) and use the resulting model to test the hypothesis $H_0 : h_i = 0$. We can also directly interpret $\tau(x) = \mu_1(x) - \mu_0(x)$ as CARE(x), and we can recover the ERPF from our estimate of π.

We estimate the model via a stagewise procedure. First, we fit a nonparametric regression function $\hat{\mu}_0(x)$ to the expected outcomes for the untreated population. Next, we use a nonparametric density estimator to marginally model the distribution $\hat{g}$ of the residuals $y - \hat{\mu}_0(x)$ on the untreated population. Finally, we use an EM algorithm on the treated population to fit nonparametric regression functions $\hat{\pi}$ and $\hat{\mu}_1$ for the prior and responder mean, with $\hat{\mu}_0(x)$ and $\hat{g}$ fixed. Implementation details are deferred to the appendix. Given these estimates, we can apply the selection procedure from Section 2.2.

While FDR control is guaranteed if we have the true posteriors [11], in practice we will incur some estimation error with finite samples. Under some regularity assumptions, we can show that the estimation errors smoothly translate to both FDR and power, where the power of a procedure that selects a subset $V \subset \{i \in [n] : t_i = 1\}$ is the expected fraction of treated individuals with $h_i = 1$ that are included in V.

Theorem 5. *Let $\epsilon, \lambda, L, U > 0$ be given. Suppose for $i = 1, \ldots, n$: (i) g is λ-Lipschitz and bounded above by U, (ii) $(1 - \pi(x_i))g(y_i - \mu_0(x_i)) + \pi(x_i)g(y_i - \mu_1(x_i)) \geq L$, (iii) $|\pi(x_i) - \hat{\pi}(x_i)|, |\mu_0(x_i) - \hat{\mu}_0(x_i)|, |\mu_1(x_i) - \hat{\mu}_1(x_i)| \leq \epsilon$, and (iv) $|g(y) - \hat{g}(y)| \leq \epsilon$ for all $y \in \mathbb{R}$. If $\epsilon \leq \min\{1, L/4(U + 2(\lambda + 1))\}$, then the additive model procedure has FDR bounded above by $\alpha+\delta$ and power bounded below by $\frac{1-\alpha-\delta}{1-\alpha+\delta+1/(n_{\mathrm{OPT}}(\alpha-\delta)+1)} \mathrm{OPT}(\alpha-\delta)$, where $\delta = \frac{2U(U+2(\lambda+1))}{L^2}\epsilon$, $\mathrm{OPT}(\beta)$ is the power of the Bayes-optimal procedure achieving FDR $\leq \beta$, and $n_{\mathrm{OPT}}(\beta)$ is the corresponding number of selections.*

Assumptions (i) and (ii) require the residual distribution and the treatment distributions to be well-behaved and are similar to standard assumptions made in the density estimation literature, while assumptions (iii) and (iv) reflect the quality of the estimators.

It is worth noting that the number of selections plays a role in the power approximation due to the discrete nature of the selection problem. However, as the number of optimal selections increases, the approximation ratio tends to $(1 - \alpha - \delta)/(1 - \alpha + \delta)$. Thus, so long as we are reasonably accurate in our estimates, the realized FDR will be close to the target and the power will tend towards optimality.

4 Causal two-groups with nonparametric models

The additivity assumption in Equation (4) is often violated in practice. For instance, Efron and Feldman [12] observed that higher rates of compliance were associated with increased variability in outcomes for both the treatment and control groups, and that the increase in variance differed between groups, suggesting that the error models for f_0 and f_1 were different. Here, we propose a generalized empirical Bayes procedure for estimation in the causal two-groups model. The goal is to be maximally flexible to the true data-generating model. As such, we remove the assumption of additive errors and instead directly model the outcome distributions in Equation (1). The starting point for this approach is the following result.

Lemma 6. *Suppose the generative model from Equation* (1) *induces treated distribution $f_t(y \mid x)$ and untreated distribution $f_0(y \mid x)$. Then for any x, $\pi(x) \geq \pi^\star(x) := 1 - \min_y \frac{f_t(y|x)}{f_0(y|x)}$. Moreover, for any $\pi(x) \in [\pi^\star(x), 1]$, there exists a responder distribution f_1^π such that $f_t(y \mid x) = (1-\pi(x))f_0(y \mid x) + \pi(x)f_1^\pi(y \mid x)$.*

The implication of Lemma 6 is that given the observed untreated and treated distributions, one can compute the most conservative mixing function, $\pi^\star$, by minimizing their ratio. This leads to a valid test for an observation (x, y) that controls the Type I error rate:

$$\hat{h}_{\text{Bayes}} = \mathbb{1}\left[\frac{(1 - \pi^\star(x))f_0(y \mid x)}{f_t(y \mid x)} \leq \alpha\right]. \tag{5}$$

Thus, empirical Bayes testing in the fully general causal two-groups model can be reduced to finding suitable estimates of f_0 and f_t.

Estimating the non-response and treatment distributions. We take a nonparametric approach to approximating f_0 and f_t, using a k-nearest-neighbor version of the Rosenblatt kernel conditional density estimator [19, 42]:

$$\hat{f}_t(y \mid x) = \frac{\sum_{i=1}^k K_{h_1}(x \mid x_{j_{x,1,i}})K_{h_2}(y \mid y_{j_{x,1,i}})}{\sum_{i=1}^k K_{h_1}(x \mid x_{j_{x,1,i}})},$$

where K_h is a probability kernel with bandwidth $h > 0$ and $j_{x,t,i}$ indexes the i-th nearest neighbor of x with treatment t. $\hat{f}_0$ is defined analogously. The parameters h_1, h_2, and k are selected via leave-one-out cross-validation.

Selecting responders. The test in Equation (5) relies on ratios of densities, making it sensitive to errors in our estimates. To address this issue, we use bootstrap sampling to create a population of conditional densities and estimate the ratios by using upper/lower quantiles to produce conservative ratio estimates. The final null posterior probability estimate for observation (x_i, y_i) is given by

$$\hat{w}_i = \frac{\hat{f}_{0,1-q}(y_i \mid x_i)}{\hat{f}_{t,q}(y_i \mid x_i)} \cdot \min_y \frac{\hat{f}_{t,1-q}(y \mid x_i)}{\hat{f}_{0,q}(y \mid x_i)}, \tag{6}$$

where $\hat{f}_{t,q}(y \mid x)$ is the q-th quantile (under the bootstrap distribution) of the conditional treatment density of y given x, and $\hat{f}_{0,q}(y \mid x)$ is defined analogously for the untreated distribution. As before, we select the largest subset of samples whose average $\hat{w}_i$ is at most α. The following result, analogous to Theorem 5, shows that when the conditional distributions are estimated accurately enough, the procedure results in proper FDR control and near-optimal power.

Theorem 7. *Let $\epsilon, L, U > 0$. Suppose for all $i = 1, \ldots, n$ and $y \in \mathbb{R}$: (i) $L \leq f_t(y \mid x_i), f_0(y \mid x_i) \leq U$, (ii) $|f_t(y \mid x_i) - \hat{f}_t(y \mid x_i)| \leq \epsilon$, and (iii) $|f_0(y \mid x_i) - \hat{f}_0(y \mid x_i)| \leq \epsilon$. If $\epsilon \leq \min(1, L/2)$, then FDR $\leq \alpha + \delta$ where $\delta = \frac{8U}{L^2}(1 + \frac{12U}{L^2})\epsilon$, and power $\geq \frac{1-\alpha-\delta}{1-\alpha+\delta+1/(n_{\mathrm{OPT}}(\alpha-\delta)+1)} \mathrm{OPT}(\alpha - \delta)$, where OPT and n_{OPT} are as in Theorem 5 but w.r.t. the Bayes-optimal procedure under $\pi^\star$.*

We can further control FDR with an empirical estimate, calculated by performing the above selection procedure on both the treated and untreated populations over a grid of nominal FDR values $\alpha_1 < \cdots < \alpha_m$, and taking e_k to be the ratio of the number of untreated selected at level α_k over the number of treated selected at level α_k. Selection at level α proceeds by finding the largest $\alpha_k \leq \alpha$ such that $e_k \leq \alpha_k$. We call this procedure *empirical control*.

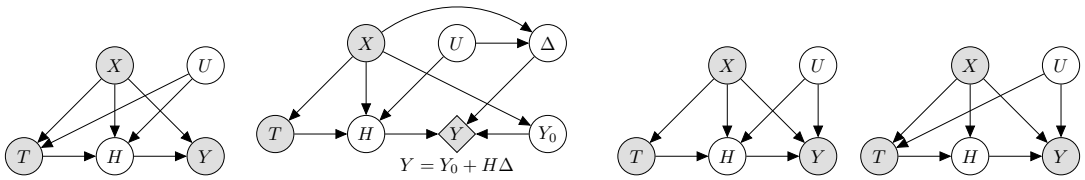
Estimand intervals. Beyond providing a test with valid FDR, Lemma 6 also provides insights into the range of effects that can be ascribed to a possible responder distribution. Specifically, even though one cannot estimate a latent response effect from an unidentifiable model, there is a non-trivial *interval* of feasible effects.

Corollary 8. *Let f_0, f_t, and $\pi^\star$ be as in Lemma 6. For any valid responder distribution f_1,*

$$\mathbb{E}_{f_1}[Y|X = x] \in [\mu_t(x), \mu_1^\star(x)] \cup [\mu_1^\star(x), \mu_t(x)]\,,$$

where $\mu_t(x) = \mathbb{E}[Y|X = x, T = 1]$ and $\mu_1^\star(x) = \frac{1}{\pi^\star(x)}(\mu_t(x) - \mu_0(x)) + \mu_0(x)$. Moreover, every value in this interval is attained by some valid f_1.

Thus, for any individual x, the true CARE lies in an interval bounded on one side by the CATE ($\mu_t(x) - \mu_0(x)$) and on the other by the extremal CARE ($\mu_1^\star(x) - \mu_0(x)$), using the convention that $[a, b] = \emptyset$ when $b < a$. Moreover, the width of this interval is exactly $|\mathrm{CATE}(x)|\,(1-\pi^\star(x))/\pi^\star(x)$.



(a) Response confounding (b) Effect-size confounding (c) Total effect confounding (d) Canonical confounding

Figure 2: Four types of latent confounding in the causal two-groups model. In (b), the outcome Y consists of a baseline variable Y_0 plus an effect Δ that applies only to responders, and U influences Δ but not Y_0.

5 Robustness against unmeasured confounding

Latent confounding in a C2G model occurs when an unobserved variable U affects two or more variables in $\{T, H, Y\}$. The risk of latent confounding is that the assumptions made in a C2G model may be violated, rendering its guarantees invalid. We say a C2G model remains *well-specified* under latent confounding if marginalizing over U yields the assumed distributional structure:

$$\begin{aligned} \textstyle\int p(y|x, u, t{=}0)\, p(u|x, t = 0)\, du &= f_0(y \mid x) = p(y|x, h{=}0, t{=}1)\,, \\ \textstyle\int p(y|x, u, t{=}1)\, p(u|x, t{=}1)\, du &= (1 - \pi(x))f_0(y \mid x) + \pi(x)f_1(y \mid x)\,. \end{aligned} \tag{7}$$

If the model is well-specified, then all inferences from Sections 3 and 4 remain valid. The following result summarizes the robustness of each C2G model to the different types of unobserved confounding.

Proposition 9. *Let U be an unobserved variable that may affect any of T, H, and Y.*

Method	Additive			Nonadditive		
	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 1$	$\tau = 3$	$\tau = 5$
Frequentist	**.04±.05** / .00±.00	**.02±.01** / .22±.05	**.02±.00** / .67±.06	**.02±.00** / .38±.06	**.02±.00** / .57±.04	**.02±.00** / .65±.03
Add-C2G	**.13±.06** / .01±.01	**.11±.02** / .48±.08	**.11±.01 / .88±.03**	**.08±.02** / .45±.05	**.05±.01** / .53±.03	**.03±.01** / .54±.03
NP-C2G	**.06±.04** / .00±.00	**.05±.01** / .44±.06	**.08±.01 / .84±.05**	**.07±.01 / .51±.06**	**.09±.01 / .68±.04**	**.11±.01 / .74±.03**
NP-Oracle	**.07±.03 / .04±.01**	**.10±.01 / .67±.05**	**.10±.00 / .92±.03**	**.09±.01 / .58±.05**	**.09±.01 / .72±.03**	**.12±.03 / .79±.03**

Table 1: FDR / Power on synthetic data ($N = 1000$, $\alpha = 0.1$). $\pm$ denotes 95% confidence intervals. **Bolded** FDR indicates valid control; **bolded** power indicates highest (up to 95% CI).

(i) The nonparametric C2G model is well-specified if and only if $Y \perp\!\!\!\perp T \mid X, H{=}0$.

(ii) The additive C2G model is well-specified if and only if, in addition to (i), the responder distribution is a location shift of the untreated distribution: $p(y|T = 1, H = 1, x) = g(y - \mu_1(x)) = p(y + \mu_0(x) - \mu_1(x)|T = 0, x)$ *for all* x.

Moreover, (i) holds if $Y \perp\!\!\!\perp T|X, U, H$ *and either (a)* $Y \perp\!\!\!\perp U \mid X, H{=}0$ *or (b)* $U \perp\!\!\!\perp (T, H) \mid X$.

As a consequence of Proposition 9, we can break down whether or not the C2G models are robust to the confounding scenarios depicted in Figure 2.

Corollary 10. *Under the scenarios in Figure 2, we have the following results.*

(a) Under response confounding, the nonparametric C2G model is always well-specified, and the additive C2G model is well-specified whenever the outcome distributions $p(y \mid x, h)$ *have the additive form of Equation* (4)*.*

(b) Under effect-size confounding, the nonparametric C2G model is always well-specified, but the additive C2G model need not be, even when $Y_0 \mid X$ *satisfies additivity.*

(c)–(d) Under total effect or canonical confounding, neither model need be well-specified.

6 Empirical evaluations

Synthetic simulations. We compare the additive (Add-C2G) and nonparametric (NP-C2G) C2G procedures against a frequentist baseline (Benjamini-Hochberg-corrected p-values from the NP-C2G untreated model) and an oracle baseline (NP-C2G with known responder/non-responder distributions). Appendix F contains additional comparisons against FDR regression [45], BART [18], and causal forests [53], and Appendix G contains semi-synthetic drug response simulations. In the synthetic settings considered here, $X \sim \mathcal{N}(0, I_{10}/10)$, $H|T{=}1$ follows a random logistic regression, and the effect size is controlled by a parameter $\tau \in \{1, 3, 5\}$. In the additive setting, outcome distributions are normal with means that are random functions of X. In the nonadditive setting, outcomes additionally depend on exogenous random variables. Each simulation uses 50 random seeds.

Table 1 shows results for $N = 1000$. All methods generally control FDR, but the frequentist baseline sacrifices substantial power. Both C2G methods achieve high power, and NP-C2G is competitive with the oracle in the nonadditive setting. We also evaluate the ARE interval estimates from Corollary 8 using the Jaccard index against oracle intervals and see that larger effect sizes and more data yield more accurate intervals (see Appendix F).

Outlier detection. The C2G framework naturally applies to outlier detection, where a set of known inliers corresponds to the untreated population, a contaminated test set corresponds to the treated population, and outliers within the test set correspond to responders. In this setting, there are no covariates X, but the outcome variable Y is possibly high-dimensional. Equation (1) simplifies to

$$y \mid h \sim f_h(y), \quad h \mid t{=}0 = 0, \quad h \mid t{=}1 \sim \text{Bern}(\pi), \quad t \sim \text{Bern}(\phi), \tag{8}$$

where π and ϕ are constants. We can apply Lemma 6 with constant x to reduce the test in Equation (5) to the problem of estimating density ratios $f_t(y)/f_0(y)$. Rather than estimating the densities separately, we learn these ratios directly via a binary classification approach. After training a random forest to predict T from Y, we apply Bayes' rule:

$$\frac{P(T{=}0)}{P(T{=}1)} \cdot \frac{P(T{=}1 \mid Y{=}y)}{P(T{=}0 \mid Y{=}y)} = \frac{f_t(y)}{f_0(y)}.$$

Because density ratios are difficult to estimate in finite samples, we form conservative estimates using the quantiles of the random forest ensemble. For decision tree j in an ensemble of m trees, let $g_j(y)$ denote its estimate of $P(T{=}1 \mid Y{=}y)$. For a given quantile q, we define

$$\hat{r}_q(y) = \frac{\#\{i : t_i{=}0\}}{\#\{i : t_i{=}1\}} \text{ quantile}\left(\left(\frac{g_j(y)}{1-g_j(y)}\right)_{j=1}^{m},\ q\right),$$

and form the null posterior estimate $\hat{w}_i = \frac{1}{\hat{r}_q(y_i)}(1 - \min_j \hat{r}_q(y_j))$, and then select as in Section 4. Smaller values of q lead to more conservative selections.

We compare against AdaDetect [33], a state-of-the-art method for this setting, on three benchmark datasets (`credit-card`, `mammography`, `shuttle`). For each dataset, both methods receive the same 500 inliers as the null set, and we construct the test set from 400 inliers and 100 outliers (30 random seeds). Table 2 shows that both methods control FDR, but NP-C2G substantially outperforms AdaDetect in power, particularly at $\alpha = 0.01$ where AdaDetect has zero power across all datasets.

7 Case study: Immune checkpoint inhibitors

We apply the NP-C2G model to a dataset tracking survival of cancer patients treated with immune checkpoint inhibitors [44]. Each patient had targeted next-generation sequencing performed on their tumors, recording the mutational state of 341–468 genes. Survival information was converted to numerical outcomes via a Cox proportional hazards model (see Appendix H for details).

Patients received one of three treatments: anti-CTLA4, anti-PD-1/PD-L1, or a combination of both. We consider patients receiving anti-CTLA4 ($n = 99$) as the baseline group and those receiving combination therapy ($n = 255$) as the treatment group. A responder is then a patient who received a benefit from PD-1/PD-L1 therapy beyond the baseline anti-CTLA4 effect. This stratification is clinically valuable because immunotherapies can trigger life-threatening complications, and identifying which patients benefit from combination therapy could inform future treatment decisions.

We fit NP-C2G as outlined in Section 4. For a fine-grained grid of nominal FDR thresholds $\alpha \in (0, 0.25]$, we calculate the rejection set S_α. To detect biomarkers of response, we compute Fisher's exact test for each gene on the contingency table of mutation status versus membership in S_α, testing whether the mutation is over-represented among responders (*more favorable*) or under-represented (*less favorable*). We apply Benjamini-Hochberg correction and record genes with $q < 0.1$. For each detected biomarker, we report the CARE interval (Corollary 8, averaged over carriers) and the conditional ERPF (CERPF) using the conservative prior $\pi^\star$.

Table 3 displays results for the selected genes. Among more favorable mutations, VHL has the smallest rejection threshold, smallest q-value, largest CERPF, and an entirely positive CARE interval. This is consistent with its role as a positive biomarker for anti-PD-1/PD-L1 therapies [25]. SETD2 mutations have also been linked to favorable immunotherapy outcomes [22, 30]. Among less favorable mutations, BRAF has the strongest signal and has been implicated in increased toxicity of combination anti-CTLA4/anti-PD-1 treatments [38], while TERT mutations are associated with less favorable anti-PD-1 outcomes [29]. The results for ERBB4 and GRIN2A represent potential novel discoveries. Given their small CERPF, it is likely they only affect a small subset of patients and may warrant further investigation to find more fine-grained molecular markers of response.

	`credit-card`		`shuttle`		`mammography`	
α	FDR	Power	FDR	Power	FDR	Power
AdaDetect						
0.01	**0.0±0.0**	0.0±0.0	**0.0±0.0**	0.0±0.0	**0.0±0.0**	0.0±0.0
0.1	**.05±.02**	**.84±.01**	**.05±.02**	.38±.10	**.08±.02**	**.98±.02**
NP-C2G						
0.01	**0.0±.002**	**.75±.07**	**.01±.01**	**.19±.08**	**0.0±.002**	**.66±.13**
0.1	**.08±.01**	**.84±.01**	**.08±.01**	**.61±.03**	**.07±.02**	**.96±.02**

Table 2: Outlier detection results. $\pm$ denotes 95% confidence intervals. **Bolded** indicates valid FDR/highest power (up to 95% confidence intervals).

	More favorable		Less favorable			
Mutation	SETD2	VHL	BRAF	ERBB4	GRIN2A	TERT
Smallest α	0.07	0.029	0.149	0.239	0.247	0.201
Largest -$\log_{10}(q)$	1.577	7.511	3.373	1.125	1.003	1.963
CARE interval	(-4.3, 1.4)	(1.3, 2.1)	(-8.5, 0.2)	(-9.0, 1.6)	(-7.8, 1.1)	(-7.2, 1.2)
CERPF	0.39	0.62	0.22	0.16	0.15	0.25

Table 3: *Smallest* α denotes the smallest nominal FDR level α of the NP-C2G procedure that results in the gene being selected. *Largest* $-\log_{10}(q)$ denotes the largest $-\log_{10}(q)$-value observed for the selected gene. *CARE interval* denotes the average of the treatment effect intervals produced by NP-C2G over the population with the given mutation. *CERPF* denotes the average of the lower-bound prior probabilities produced by NP-C2G over the treated population with the given mutation.

8 Discussion

The emergence of large, real-world electronic health record datasets [e.g., 22] presents both an opportunity and a challenge for precision medicine. Cohorts of tens of thousands of patients enable detection of significant associations in fine-grained subgroups, but these observational data are susceptible to the usual pitfalls of causal inference. The methods presented here represent a step towards new methodology for this modern setting, but many interesting questions still remain open.

- *Identifiability characterization.* We showed identifiability for parametric and additive semi-parametric models, but a complete characterization of what modeling assumptions ensure identifiability remains open.
- *Non-binary treatments.* The treatments considered in this paper are binary, but many settings involve non-binary treatments. For example, drugs have doses, treatments have schedules, and training programs span multiple sessions. Extending the C2G framework to continuous or multi-valued treatments is a natural next step.
- *Relaxing i.i.d. assumptions.* Our FDR control results (Theorem 5 and Theorem 7) only require accurate density estimates, not i.i.d. observations, but our estimation procedures do rely on i.i.d. structure. Developing algorithms that can tolerate interference or other forms of dependence is an important direction.

Finally, our immunotherapy case study presents a potentially new way to analyze combination therapy studies. Rather than simply considering the difference in survival time, one can now estimate how many patients actually benefit from combination therapy. Application to datasets of different treatment regimens may reveal that certain combinations are effective only for a small group of patients who have an outsized effect, suggesting the need for better patient stratification.

Limitations. The guarantees for the algorithms presented in this paper depend on the assumption of Equation (1) that treated non-responders have the same distribution over outcomes as untreated individuals, conditioned on the covariates X. Ultimately, this assumption cannot be checked from observable data, and it can be violated under certain types of latent confounding, as shown in Section 5. Additionally, the causal interpretation of the C2G model in Section 2.4 rests on further assumptions that often cannot be verified and indeed have been shown to be violated in certain cases [12]. Finally, the algorithms presented here require accurate estimates of various functions, which become harder to obtain as the dimensionality of the covariate space grows.

Acknowledgements

We gratefully acknowledge support by the NIH/NCI (R37 CA271186, U54 CA274492, P30 CA008748), Break Through Cancer, and the Maurice Campbell Initiative at Memorial Sloan Kettering Cancer Center.

References

[1] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[2] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[3] I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 16434–16445, 2020.

[4] J. A. Boatman, D. M. Vock, J. S. Koopmeiners, and E. C. Donny. Estimating causal effects from a randomized clinical trial when noncompliance is measured with error. *Biostatistics*, 19 (1):103–118, 2017.

[5] X. Chen, D. G. Robinson, and J. D. Storey. The functional false discovery rate with applications to genomics. *Biostatistics*, 22(1):68–81, 2019.

[6] H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[7] G. da Cunha Santos, F. A. Shepherd, and M. S. Tsao. EGFR mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, 6(1):49–69, 2011.

[8] B. Duan, L. Wasserman, and A. Ramdas. Interactive identification of individuals with positive treatment effect while controlling false discoveries. *Journal of Causal Inference*, 12(1), 2024.

[9] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[10] B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.

[11] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press, 2012.

[12] B. Efron and D. Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

[13] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[14] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1): 21–29, 2002.

[15] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[16] P. R. Hahn, J. S. Murray, and C. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523*, 2017.

[17] M. Hidrobo, J. Hoddinott, A. Peterman, A. Margolies, and V. Moreira. Cash, food, or vouchers? evidence from a randomized experiment in northern ecuador. *Journal of Development Economics*, 107:144–156, 2014.

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

[19] M. P. Holmes, A. G. Gray, and C. L. Isbell Jr. Fast nonparametric conditional density estimation. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 175–182, 2007.

[20] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577, 2016.

[21] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, and M. J. Garnett. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740—754, 2016.

[22] J. Jee, C. Fong, K. Pichotta, T. N. Tran, A. Luthra, M. Waters, C. Fu, M. Altoe, S.-Y. Liu, S. B. Maron, M. Ahmed, S. Kim, M. Pirun, W. K. Chatila, I. de Bruijn, A. Pasha, R. Kundra, B. Gross, B. Mastrogiacomo, T. J. Aprati, D. Liu, J. Gao, M. Capelletti, K. Pekala, L. Loudon, M. Perry, C. Bandlamudi, M. Donoghue, B. A. Satravada, A. Martin, R. Shen, Y. Chen, A. R. Brannon, J. Chang, L. Braunstein, A. Li, A. Safonov, A. Stonestrom, P. Sanchez-Vela, C. Wilhelm, M. Robson, H. Scher, M. Ladanyi, J. S. Reis-Filho, D. B. Solit, D. R. Jones, D. Gomez, H. Yu, D. Chakravarty, R. Yaeger, W. Abida, W. Park, E. M. O'Reilly, J. Garcia-Aguilar, N. Socci, F. Sanchez-Vega, J. Carrot-Zhang, P. D. Stetson, R. Levine, C. M. Rudin, M. F. Berger, S. P. Shah, D. Schrag, P. Razavi, K. L. Kehl, B. T. Li, G. J. Riely, N. Schultz, A. Lisman, G. Zhao, W. K. Chatila, X. Li, A. Kohli, D. Moore, R. Lim, T. Pollard, R. Sheridan, A. Wang, C. Chennault, M. Wilson, H. Zhang, R. Pimienta, S. Rangavajhala, G. Subramanian, J. Garcia, N. Rachuri, K. Boehm, M. Parker, H. Walch, S. Nandakumar, J. Eichholz, A. Kris, P. Manca, X. Bai, T. Agbamu, J. U, X. Bi, and M. C. D. S. I. Group. Automated real-world data integration improves cancer outcome prediction. *Nature*, 636(8043):728–736, 2024.

[23] Z. Jiang and P. Ding. Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, 36(4):493–508, 2021.

[24] Z. Jiang, S. Yang, and P. Ding. Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society: Series B (Methodological)*, 84(4):1423–1445, 2022.

[25] M. Jiao, M. Hu, D. Pan, X. Liu, X. Bao, J. Kim, F. Li, and C.-Y. Li. VHL loss enhances antitumor immunity by activating the anti-viral DNA-sensing pathway. *iScience*, 27(7), 2024.

[26] F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

[27] L. Lei and W. Fithian. AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 80(4):649–679, 2018.

[28] A. Li and R. F. Barber. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 81(1):45–74, 2019.

[29] H. Li, J. Li, C. Zhang, C. Zhang, and H. Wang. TERT mutations correlate with higher TMB value and unique tumor microenvironment and may be a potential biomarker for anti-CTLA4 treatment. *Cancer Med*, 9(19):7151–7160, 2020.

[30] M. Lu, B. Zhao, M. Liu, L. Wu, Y. Li, Y. Zhai, and X. Shen. Pan-cancer analysis of SETD2 mutation and its association with the efficacy of immunotherapy. *npj Precision Oncology*, 5(1): 51, 2021.

[31] S. Lu, Z. Jiang, and P. Ding. Principal stratification with continuous post-treatment variables: Nonparametric identification and semiparametric estimation. *arXiv preprint arXiv:2309.12425*, 2023.

[32] C. F. Manski. *Partial identification of probability distributions*. Springer, 2003.

[33] A. Marandon, L. Lei, D. Mary, and E. Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

[34] R. Martin and S. T. Tokdar. Asymptotic properties of predictive recursion: Robustness and rate of convergence. *Electronic Journal of Statistics*, 3:1455–1472, 2009.

[35] R. Martin and S. T. Tokdar. Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika*, 98(3):567–582, 2011.

[36] R. McNamee. Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine*, 28(21):2639–2652, 2009.

[37] R. K. Patra and B. Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 78(4): 869–893, 2016.

[38] I. Pires da Silva, D. Zakria, T. Ahmed, C. Trojanello, F. Dimitriou, C. Allayous, C. Gerard, L. Zimmer, S. Lo, O. Michielin, C. Lebbe, J. Mangana, P. A. Ascierto, D. B. Johnson, M. Carlino, A. Menzies, and G. Long. Efficacy and safety of anti-PD1 monotherapy or in combination with

ipilimumab after BRAF/MEK inhibitors in patients with BRAF mutant metastatic melanoma. *Journal for ImmunoTherapy of Cancer*, 10(7), 2022.

[39] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

[40] S. I. Resnick. *A probability path*. Springer Science & Business Media, 2013.

[41] L. A. Rojas, Z. Sethna, K. C. Soares, C. Olcese, N. Pang, E. Patterson, J. Lihm, N. Ceglia, P. Guasp, A. Chu, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature*, 618(7963):144–150, 2023.

[42] M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate analysis, II*, pages 25–31. Academic Press, 1969.

[43] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[44] R. M. Samstein, C.-H. Lee, A. N. Shoushtari, M. D. Hellmann, R. Shen, Y. Y. Janjigian, D. A. Barron, A. Zehir, E. J. Jordan, A. Omuro, T. J. Kaley, S. M. Kendall, R. J. Motzer, A. A. Hakimi, M. H. Voss, P. Russo, J. Rosenberg, G. Iyer, B. H. Bochner, D. F. Bajorin, H. A. Al-Ahmadie, J. E. Chaft, C. M. Rudin, G. J. Riely, S. Baxi, A. L. Ho, R. J. Wong, D. G. Pfister, J. D. Wolchok, C. A. Barker, P. H. Gutin, C. W. Brennan, V. Tabar, I. K. Mellinghoff, L. M. DeAngelis, C. E. Ariyan, N. Lee, W. D. Tap, M. M. Gounder, S. P. D'Angelo, L. Saltz, Z. K. Stadler, H. I. Scher, J. Baselga, P. Razavi, C. A. Klebanoff, R. Yaeger, N. H. Segal, G. Y. Ku, R. P. DeMatteo, M. Ladanyi, N. A. Rizvi, M. F. Berger, N. Riaz, D. B. Solit, T. A. Chan, and L. G. T. Morris. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet*, 51(2):202–206, 2019.

[45] J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery-rate regression: An application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510), 2015.

[46] K. Seipel, M. Abbühl, U. Bacher, H. Nilius, M. Daskalakis, and T. Pabst. Clinical impact of single nucleotide polymorphism in CD-19 on treatment outcome in FMC63-CAR-T cell therapy. *Cancers*, 15(11):3058, 2023.

[47] H. Shen and C. G. Maki. Pharmacologic activation of p53 by small-molecule MDM2 antagonists. *Curr Pharm Des*, 17(6):560–568, 2011.

[48] C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517, 2019.

[49] I. Shrier, R. J. Steele, E. Verhagen, R. Herbert, C. A. Riddell, and J. S. Kaufman. Beyond intention to treat: What is the right question? *Clinical Trials*, 11(1):28–37, 2014.

[50] W. Tansey, O. Koyejo, R. A. Poldrack, and J. G. Scott. False discovery rate smoothing. *Journal of the American Statistical Association*, 113(523):1156–1171, 2018.

[51] W. Tansey, Y. Wang, D. M. Blei, and R. Rabadan. Black box FDR. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4874–4883, 2018.

[52] H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

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

[54] H. Wang, J. Fan, Z. Chen, H. Li, W. Liu, T. Liu, Q. Dai, Y. Wang, Z. Dong, and R. Tang. Optimal transport for treatment effect estimation. In *Advances in Neural Information Processing Systems*, pages 5404–5418, 2024.

[55] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983.

[56] F. Xia, M. J. Zhang, J. Y. Zou, and D. Tse. NeuralFDR: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, pages 1540–1549, 2017.

[57] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

[58] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41: D955—61, January 2013.

A Extended discussion on related work

This paper draws inspiration from several related areas in statistics and machine learning. Here we flesh out the connected areas more fully.

Multiple hypothesis testing with covariates. Several approaches to multiple testing incorporate covariates as side information. These methods generally work under the two-groups model [9, 10, 11] where a test statistic is drawn either from the null or alternative with some prior probability. However, both frequentist [20, 56, 27, 5, 28] and Bayesian [45, 37, 50, 51] approaches assume that the covariates only impact the likelihood of treatment having an effect. This is often a reasonable assumption in laboratory experiments, where biological replicates can be tested repeatedly under different experimental conditions. In observational data, the assumption is unlikely to hold as covariates are often confounders that affect the probability of treatment assignment, treatment compliance, and the outcomes themselves. Figure 1a shows the graphical model for existing two-groups models.

In a similar vein to the latent response model considered here, Duan et al. [8] propose an interactive protocol to the testing problem, in which an analyst is given the covariates and outcomes but is blinded to the treatment assignments, and they must cooperate with an algorithm that knows the treatment assignments to make discoveries. In their setup, however, the blinding to treatment assignments is not intrinsic to the problem but is rather a technique to increase power without losing FDR control.

Causal inference with noncompliance. In the statistics literature, methods for handling noncompliance were originally motivated by clinical trial data. Efron and Feldman [12] observed that patients receiving placebo treatments had better outcomes as a function of compliance rate to the placebo. Frangakis and Rubin [14] generalized this in the *principal stratification* framework, with treatment assignment and compliance jointly representing potential interventions. Figure 1b shows the graphical model for a randomized controlled trial (RCT) in the principal stratification setting. These models assume compliance is observed, leading to identifiability of the causal effects; here we consider the case where compliance is entirely latent.

Other methods consider noisy or latent compliance with side information. Angrist et al. [1] introduced an instrumental variable approach to estimating average treatment effects with noisy compliance. Boatman et al. [4] consider the scenario where a noisy proxy variable is available for compliance, such as blood testing for nicotine levels when patients may be misreporting their smoking habits. Further, they assume the true parametric model is known and identifiable. Figure 1c shows the graphical model for the proxy variable case. More recent work leverages auxiliary variables combined with conditional ignorability assumptions or partial observability of the compliance variable [23, 24, 31]. These models obtain identifiability by leveraging additional information to disentangle the confounding from the causal effects. By contrast, we consider the challenging case where no proxy variables are available.

Estimating individual treatment effects in observational data. A large body of recent work in machine learning has considered the problem of predicting the conditional average treatment effect (CATE) [e.g. 48, 26, 3, 54]. In the presence of latent response, the CATE is generally an underestimate of the true effect if one could intervene, such as requiring compliance. CATE estimates also do not directly help identify which samples may have responded to treatment, as they integrate out the probability of response in a given sample. However, methods that provide uncertainty quantification can theoretically be used to conduct hypothesis tests for whether a given treatment has any effect on a given individual. Causal forests [53] and Bayesian additive regression trees (BART) [6, 18] are two popular methods capable of providing such confidence intervals of causal effects. Synthetic and semi-synthetic simulations in Appendices F and G show that confidence intervals from these models are not sufficient to overcome the fact that they do not incorporate explicit notions of null effects among nonresponders when it comes to identifying individuals that responded to treatment.

Partial identification. The most general form of the causal two-groups model can be seen as falling inside the partial indentifiability framework [32], in which one creates estimates of a non-identifiable parameter or distribution by estimating a tight set that it must belong to. In our nonparametric C2G setting, we give an explicit form for this set, and we further show that it can be used to construct tests with proper FDR control and to create intervals containing conditional latent response effects. In a similar vein, [2] derived tight intervals for average treatment effects (ATE) under noncompliance

(Figure 1d), where compliance is observed but there is an unobserved confounder of both compliance and response.

B Additional results and details from Section 2

B.1 Proof of Proposition 1

Consider the following densities over $[0, \infty)$:

$$\begin{aligned} f_0(y) &= \frac{3}{2}e^{-y} - e^{-2y} \\ f_1^{(1)}(y) &= e^{-y} \\ f_1^{(2)}(y) &= 2e^{-2y}. \end{aligned}$$

It can be readily verified that these are all valid probability densities. Moreover, we can see that f_0 is log-concave over its domain since

$$\frac{d^2}{dy^2} \log f_0(y) = -\frac{2e^y}{3(\frac{2}{3} - e^y)^2} < 0 \text{ for all } y \geq 0.$$

$f_1^{(1)}$ and $f_1^{(2)}$ are also clearly log-concave since their log-densities are linear.

Suppose that f_0 is the observed null distribution. Letting $\pi_1 = \frac{3}{4}$ and $\pi_2 = \frac{1}{4}$, consider the following two observed treatment distributions:

$$\begin{aligned} f_t^{(1)}(y) &= (1 - \pi_1) f_0(y) + \pi_1 f_1^{(1)}(y) \\ f_t^{(2)}(y) &= (1 - \pi_2) f_0(y) + \pi_2 f_1^{(2)}(y). \end{aligned}$$

Using these definitions, we have

$$\begin{aligned} f_t^{(1)}(y) &= (1 - \pi_1) f_0(y) + \pi_1 f_1^{(1)}(y) \\ &= \frac{1}{4}\left(\frac{3}{2}e^{-y} - e^{-2y}\right) + \frac{3}{4}e^{-y} \\ &= \frac{9}{8}e^{-y} - \frac{1}{4}e^{-2y} \\ &= \frac{3}{4}\left(\frac{3}{2}e^{-y} - e^{-2y}\right) + \frac{1}{4} \cdot 2e^{-y} \\ &= (1 - \pi_2) f_0(y) + \pi_2 f_1^{(2)}(y) \\ &= f_t^{(2)}(y). \end{aligned}$$

Thus, we can conclude that the C2G model is unidentifiable even when both the null distribution and the alternative distribution are restricted to being log-concave.

B.2 Proof of Proposition 2

Fix x. In Equation (1), we observe distributions $f_0(y \mid x)$ and $f_t(y \mid x)$. Suppose we are given $\mu_0 = \mu_0(x)$, $\mu_1 = \mu_1(x)$, $\mu_2 = \mu_2(x)$, $\sigma_0^2 = \sigma_0^2(x)$, $\sigma_1^2 = \sigma_1^2(x)$, $\sigma_2^2 = \sigma_2^2(x)$, $\pi_1 = \pi_1(x)$, and $\pi_2 = \pi_2(x)$, where $\pi_1, \pi_2 \in (0, 1)$. By translating and scaling the space, we can assume without loss of generality that $\mu_0 = 0$ and $\sigma_0^2 = 1$. Our goal is to show that if

$$(1 - \pi_1)\mathcal{N}(y \mid 0, 1) + \pi_1 \mathcal{N}(y \mid \mu_1, \sigma_1^2) = (1 - \pi_2)\mathcal{N}(y \mid 0, 1) + \pi_2 \mathcal{N}(y \mid \mu_2, \sigma_2^2) \tag{9}$$

for all $y \in \mathbb{R}$, then we must have $\pi_1 = \pi_2$, $\mu_1 = \mu_2$, and $\sigma_1^2 = \sigma_2^2$. There are two cases to consider here.

Case 1: $\pi_1 = \pi_2$. In this case, we immediately have

$$\mathcal{N}(y \mid \mu_1, \sigma_1^2) = \mathcal{N}(y \mid \mu_2, \sigma_2^2)$$

for all $y \in \mathbb{R}$. Taking the first moment of both sides, we have $\mu_1 = \mu_2$. Conditioned on this fact, we can calculate the variance of both sides to conclude $\sigma_1^2 = \sigma_2^2$.

Case 2: $\pi_1 \neq \pi_2$**.** In this case, we will come to a contradiction. Assume without loss of generality that $\pi_1 > \pi_2$. Letting $\beta = \frac{\pi_2}{\pi_1 - \pi_2} > 0$, we can rearrange Equation (9) to get

$$\mathcal{N}(y \mid 0, 1) = (1+\beta)\mathcal{N}(y \mid \mu_1, \sigma_1^2) - \beta\mathcal{N}(y \mid \mu_2, \sigma_2^2).$$

Multiplying both sides by e^{ty} and integrating $y \in \mathbb{R}$, the moment generating function of the normal distribution gives us

$$\exp\left(t^2/2\right) = (1+\beta)\exp\left(\mu_1 t + \sigma_1^2 t^2/2\right) - \beta\exp\left(\mu_2 t + \sigma_2^2 t^2/2\right) \tag{10}$$

for all $t \in \mathbb{R}$. We will analyze Equation (10) in the limit as $t \to \infty$.

Observe first that we must have $\sigma_2^2 \leq \sigma_1^2$, since otherwise as $t \to \infty$, the RHS of Equation (10) tends to $-\infty$ while the LHS is positive.

Now consider the case where $\sigma_2^2 < \sigma_1^2$. Multiplying Equation (10) through by $e^{-\mu_1 t - \sigma_1^2 t^2/2}$, we have

$$\exp\left(-\mu_1 t - (\sigma_1^2 - 1)t^2/2\right) = 1 + \beta - \beta\exp\left((\mu_2 - \mu_1)t - (\sigma_1^2 - \sigma_2^2)t^2/2\right).$$

As $t \to \infty$, the RHS tends to $1 + \beta$. On the other hand, the LHS must tend to either 0 or $+\infty$. Thus we must have $\sigma_1^2 = \sigma_2^2$. In which case, we can rearrange Equation (10) to obtain

$$1 = \left[(1+\beta)\exp(\mu_1 t) - \beta\exp(\mu_2 t)\right]\exp\left((\sigma_1^2 - 1)t^2/2\right).$$

If $\sigma_1^2 < 1$, then the RHS must tend to 0. If $\sigma_1^2 > 1$, then the RHS must tend to $\pm\infty$. Thus, we can conclude $\sigma_1^2 = 1$, and the above equation becomes

$$1 = \left[(1+\beta)\exp(\mu_1 t) - \beta\exp(\mu_2 t)\right].$$

The only way for this to hold for all t is if $\mu_1 = \mu_2 = 0$.

B.3 Proof of Theorem 3

By our assumption that $H(0) = 0$ (which implies $H = T \cdot H(1)$), the event $\{H = 1\}$ is equal to $\{T = 1, H(1) = 1\}$. Thus, by (A1) and consistency (i.e., $Y = Y(T)$), we have

$$\mathbb{E}[Y \mid T{=}1, H{=}1, x] = \mathbb{E}[Y(1) \mid T{=}1, H(1){=}1, x] = \mathbb{E}[Y(1) \mid H(1){=}1, x].$$

Chaining (A1) and (A3) and utilizing consistency again, we have

$$\mathbb{E}[Y \mid T{=}0, x] = \mathbb{E}[Y(0) \mid x] = \mathbb{E}[Y(0) \mid H(1){=}1, x].$$

Putting it all together,

$$\text{CARE}(x) = \mathbb{E}[Y \mid T{=}1, H{=}1, x] - \mathbb{E}[Y \mid T{=}0, x] = \mathbb{E}[Y(1) - Y(0) \mid H(1){=}1, x].$$

The equality statement of ERPF follows directly from the fact that $\{H = 1\}$ equals $\{T = 1, H(1) = 1\}$, which gives us

$$\text{ERPF} = P(H{=}1 \mid T{=}1) = P(H(1){=}1 \mid T{=}1).$$

The lower bound follows from (A2), which gives us the contrapositive statement $\{Y(1) \neq Y(0)\} \subseteq \{H(1) = 1\}$, from which we get

$$P(H(1){=}1 \mid T{=}1) \geq P(Y(1) \neq Y(0) \mid T{=}1).$$

C Additional results and details from Section 3

C.1 Semi-parametric model fitting details

Here we fill in the missing details of fitting our additive causal two-groups model.

Estimating the non-response distribution. We model $\hat{\mu}_0$ using kernel ridge regression with a radial basis function kernel, and we fit it on the untreated population, using generalized cross-validation to select the bandwidth and regularization parameters [15]. After fitting $\hat{\mu}_0$, we compute the leave-one-out predictions for each point in the untreated population, which can be calculated efficiently in closed form. We use these conditional expectations to estimate the residual noise distribution g. Cross-validation generally produces a biased-upward estimate of the true error distribution [13]. By overestimating the tails of the residual distribution, we conservatively bias the posterior probability of a response downward and the local FDR estimates upward.

We estimate the distribution of residuals $\hat{g}$ nonparametrically via predictive recursion [34], which has demonstrated good rates of convergence and strong performance in other two-groups models [45, 50]. We choose the bandwidth by maximizing the predictive recursion marginal likelihood (PRML) [35]. This yields a tight fit to the data and makes the density estimation routine fully auto-tuned and data-adaptive.

Estimating the prior and response distributions. With $\hat{g}$ and $\hat{\mu}_0$ in hand, we form an estimate of the alternative distribution for the treated population. Specifically, we model each member of the treatment population as arising from the mixture model

$$(y \mid x, t = 1) \sim (1 - \pi(x))\hat{g}(y - \hat{\mu}_0(x)) + \pi(x)\hat{g}(y - \mu_1(x)) \,. \tag{11}$$

We fit $\hat{\mu}_1$ and $\hat{\pi}$ to maximize the likelihood of Equation (11) via expectation-maximization. Both $\hat{\mu}_1$ and $\hat{\pi}$ (in logit-space) are encoded as linear models on top of random Fourier features, which approximate the full kernelized models and can be fit using gradient-based methods [39]. The bandwidth parameters are selected using k-fold cross validation, and final predictions for the treated population are made on each held-out fold.

E-step. Fix the estimates $\hat{\pi}$ and $\hat{\mu}_1$. Calculate the expected posterior probability of each data point coming from $\hat{g}(y - \mu_0(x))$,

$$\hat{w}_i = \frac{(1 - \hat{\pi}(x_i))\hat{g}(y_i - \hat{\mu}_0(x))}{(1 - \hat{\pi}(x_i))\hat{g}(y_i - \hat{\mu}_0(x)) + \hat{\pi}(x_i)\hat{g}(y_i - \hat{\mu}_1(x))} \,.$$

The posterior expectations $\hat{w}$ then serve as weights in the M-step.

M-step. Fixing the weights $\hat{w}$, the optimization problem is separable in μ_1 and π,

$$\hat{\pi} = \operatorname*{argmax}_{\pi} \sum_{i=1}^{n} \left[\hat{w}_i \log(1 - \pi(x_i)) + (1 - \hat{w}_i) \log(\pi(x_i))\right]$$

$$\hat{\mu}_1 = \operatorname*{argmax}_{\mu_1} \sum_{i=1}^{n} (1 - \hat{w}_i) \log(\hat{g}(y_i - \mu_1(x_i))) \,.$$

Convergence of the EM procedure. Since $\hat{g}$ and $\hat{\mu}_0$ are fixed ahead of the EM algorithm, the only parameters that are updated by the EM procedure are $\hat{\pi}$ and $\hat{\mu}_1$. In the M-step, the π-update is a weighted logistic regression, which is concave in the coefficients of the random Fourier features, and the μ_1-update corresponds to maximizing a weighted log-likelihood. If the relevant estimation and optimization steps are sufficiently accurate, then the observed-data likelihood will be non-decreasing, and the procedure will converge to a stationary point under certain regularity conditions [55].

C.2 Proof of Theorem 4

For any x, observe that we can shift the space by $\mu_0(x)$ so that the model in Equation (4) becomes

$$f_0(y \mid x) = g(y)$$
$$f_1(y \mid x) = g(y - \tau(x)).$$

In this case, the observable distributions are the null distribution $f_0 = g$ and the treatment distribution

$$f_t(y \mid x) = (1 - \pi(x))g(y) + \pi(x)g(y - \tau(x)).$$

Fixing a particular x, the identifiability of Equation (4) can be restated as follows.

Theorem 11 (Restatement of Theorem 4)**.** *Let g denote a probability density function over $\mathbb{R}$. For $\pi \in (0,1)$, $\tau \in \mathbb{R}$, define the density function*

$$f(y;\pi,\tau) = (1-\pi)g(y) + \pi g(y-\tau).$$

Then the family of densities $\{f(y;\pi,\tau) \mid \pi \in (0,1), \tau \in \mathbb{R}\setminus\{0\}\}$ is identifiable.

We begin by recalling the characteristic function of a probability distribution. If g is a probability distribution over $\mathbb{R}$, then its characteristic function is given by

$$\varphi_g(\omega) = \mathbb{E}_{X\sim g}\left[e^{-i\omega X}\right] = \int_{-\infty}^{\infty} g(x)e^{-i\omega x}\,dx,$$

where $\omega \in \mathbb{R}$ and the integral is only well-defined when g is a probability density function. We will make use of the following elementary property of the characteristic function.

Lemma 12. *For any probability density function g, there exists a value $\omega_0 > 0$ such that $|\varphi_g(\omega)| > 0$ for all $\omega \in [0,\omega_0]$.*

Proof. This follows directly from the fact that φ_g is uniformly continuous on $\mathbb{R}$ and $\varphi_g(0) = 1$ [40]. □

With this tool in hand, we turn to the proof of Theorem 11.

Proof of Theorem 11. Let $\pi_1, \pi_2 \in (0,1)$ and $\tau_1, \tau_2 \in \mathbb{R}\setminus\{0\}$ be parameters such that $f(y;\pi_1,\tau_1) = f(y;\pi_2,\tau_2)$. Our goal is to show that we must have $\pi_1 = \pi_2$ and $\tau_1 = \tau_2$. We consider two cases.

Case 1: $\pi_1 = \pi_2$. In this case, we must have

$$\begin{aligned}
0 &= f(y;\pi_1,\tau_1) - f(y;\pi_2,\tau_2) \\
&= (1-\pi_1)g(y) + \pi_1 g(y-\tau_1) - (1-\pi_2)g(y) - \pi_2 g(y-\tau_2) \\
&= \pi_1 g(y-\tau_1) - \pi_1 g(y-\tau_2)
\end{aligned}$$

for all $y \in \mathbb{R}$. Multiplying through by $e^{i\omega y}/\pi_1$ and integrating over $\mathbb{R}$, we have

$$\begin{aligned}
0 &= \int_{-\infty}^{\infty} g(y-\tau_1)e^{i\omega y}\,dy - \int_{-\infty}^{\infty} g(y-\tau_2)e^{i\omega y}\,dy \\
&= \int_{-\infty}^{\infty} g(y)e^{i\omega(y+\tau_1)}\,dy - \int_{-\infty}^{\infty} g(y)e^{i\omega(y+\tau_2)}\,dy \\
&= \varphi_g(\omega)e^{i\omega\tau_1} - \varphi_g(\omega)e^{i\omega\tau_2},
\end{aligned}$$

which holds for all $\omega \in \mathbb{R}$. Invoking Lemma 12, there is a value a value $\omega_0 > 0$ such that $|\varphi_g(\omega)| > 0$ for all $\omega \in [0,\omega_0]$. Plugging $\omega = \min\left(\omega_0, \frac{\pi}{|\tau_1-\tau_2|}\right)$ into the above and dividing by $\varphi_g(\omega)$, we can rearrange to observe

$$e^{i\omega(\tau_1-\tau_2)} = 1.$$

This can only be true if $\omega(\tau_1-\tau_2) = 2\pi n$ for some $n \in \mathbb{Z}$. By our choice of ω, it must be the case that $\tau_1 = \tau_2$.

Case 2: $\pi_1 \neq \pi_2$. Here we will reach a contradiction. Assume without loss of generality that $\pi_1 < \pi_2$. Then we have

$$\begin{aligned}
0 &= f(y;\pi_1,\tau_1) - f(y;\pi_2,\tau_2) \\
&= (1-\pi_1)g(y) + \pi_1 g(y-\tau_1) - (1-\pi_2)g(y) - \pi_2 g(y-\tau_2) \\
&= (\pi_2-\pi_1)g(y) + \pi_1 g(y-\tau_1) - \pi_2 g(y-\tau_2)
\end{aligned}$$

for all $y \in \mathbb{R}$. Multiplying through by $e^{i\omega y}$ and integrating over $\mathbb{R}$, we have

$$\begin{aligned}
0 &= (\pi_2-\pi_1)\int_{-\infty}^{\infty} g(y)e^{i\omega y}\,dy + \pi_1\int_{-\infty}^{\infty} g(y-\tau_1)e^{i\omega y}\,dy - \pi_2\int_{-\infty}^{\infty} g(y-\tau_2)e^{i\omega y}\,dy \\
&= (\pi_2-\pi_1)\varphi_g(\omega) + \pi_1\varphi_g(\omega)e^{i\omega\tau_1} - \pi_2\varphi_g(\omega)e^{i\omega\tau_2},
\end{aligned}$$

for all $\omega \in \mathbb{R}$. Invoking Lemma 12 again, we have $|\varphi_g(\omega)| > 0$ for all $\omega \in [0, \omega_0]$. Taking $\omega = \min(\omega_0, \frac{\pi}{|\tau_1|})$, we divide through by $\varphi_g(\omega)\pi_2$ and rearrange to get

$$e^{i\omega\tau_2} = 1 - \frac{\pi_1}{\pi_2} + \frac{\pi_1}{\pi_2} e^{i\omega\tau_1} = 1 - p + pe^{i\omega\tau_1},$$

where $p \in (0, 1/2)$. Taking magnitudes of both sides, we have

$$\begin{aligned} 1 = \left|e^{i\omega\tau_2}\right| &= \left|1 - p + pe^{i\omega\tau_1}\right| \\ &= |1 - p + p(\cos(\omega\tau_1) + i\sin(\omega\tau_1))| \\ &= \sqrt{(1 - p + p\cos(\omega\tau_1))^2 + p^2\sin^2(\omega\tau_1)} \\ &= \sqrt{1 - 2p(1-p)(1 - \cos(\omega\tau_1))} < 1, \end{aligned}$$

where the last line follows from our choice of ω. Thus, we have reached a contradiction. □

C.3 Proof of Theorem 5

We require the following result showing that we our posterior estimates are accurate.

Lemma 13. *Suppose that the conditions of Theorem 5 hold. Then for all $i = 1, \ldots, n$, the absolute difference*

$$\left|\frac{(1-\pi(x_i))g(y_i - \mu_0(x_i))}{(1-\pi(x_i))g(y_i - \mu_0(x_i)) + \pi(x_i)g(y_i - \mu_1(x_i))} - \frac{(1-\hat{\pi}(x_i))\hat{g}(y_i - \hat{\mu_0}(x_i))}{(1-\hat{\pi}(x_i))\hat{g}(y_i - \hat{\mu_0}(x_i)) + \hat{\pi}(x_i)\hat{g}(y_i - \hat{\mu_1}(x_i))}\right|$$

is bounded above by $\frac{2U(U+2(\lambda+1))}{L^2}\epsilon$.

Proof. Pick an index $i = 1, \ldots, n$. We use the following notation

$$g_0 = g(y_i - \mu_0(x_i)),\ g_1 = g(y_i - \mu_1(x_i)),\ \hat{g_0} = \hat{g}(y_i - \hat{\mu_0}(x_i)),\ \hat{g_1} = \hat{g}(y_i - \hat{\mu_1}(x_i)).$$

The smoothness of g combined with the approximation factor of $\hat{\mu_0}$ implies

$$|g_0 - \hat{g_0}| \leq |g(y_i - \mu_0(x_i)) - g(y_i - \hat{\mu_0}(x_i))| + |g(y_i - \hat{\mu_0}(x_i)) - \hat{g}(y_i - \hat{\mu_0}(x_i))| \leq \epsilon(\lambda + 1).$$

A symmetrical argument shows the same holds for $|g_1 - \hat{g_1}|$. Thus, we can conclude the following

$$\begin{aligned} |\pi(x_i)g_1 - \hat{\pi}(x_i)\hat{g}_1| &\leq \epsilon(U + 2(\lambda+1)) \\ |(1-\pi(x_i))g_0 - (1-\hat{\pi}(x_i))\hat{g}_0| &\leq \epsilon(U + 2(\lambda+1)) \\ (1-\hat{\pi}(x_i))\hat{g}_0 + \hat{\pi}(x_i)\hat{g}_1 &\geq L/2 \\ (1-\pi(x_i))g_0 + \pi(x_i)g_1 &\leq U. \end{aligned}$$

Using the shorthand $a = (1-\pi(x_i))g_0$, $\hat{a} = (1-\hat{\pi}(x_i))\hat{g}_0$, $b = \pi(x_i)g_1$, and $\hat{b} = \hat{\pi}(x_i)\hat{g}_1$, we have

$$\begin{aligned} \left|\frac{a}{a+b} - \frac{\hat{a}}{\hat{a}+\hat{b}}\right| &= \left|\frac{\hat{a}b - \hat{b}a}{(a+b)(\hat{a}+\hat{b})}\right| \\ &\leq \frac{2}{L^2}\left|\hat{a}b - \hat{b}a\right| \\ &\leq \frac{2}{L^2}(a+b)\epsilon(U+2) = \frac{2U(U+2(\lambda+1))}{L^2}\epsilon. \end{aligned}$$

□

Our next lemma is an elementary result concerning sums of probabilities.

Lemma 14. *Let $p_1, \hat{p}_1, p_2, \hat{p_2}, \ldots, p_n, \hat{p}_n, \alpha, \delta \in [0,1]$ such that $|p_i - \hat{p}_i| \leq \delta$ for all $i = 1, \ldots, n$. Let $V, \hat{V} \subset \{1, \ldots, n\}$ be the largest subsets such that*

$$\frac{1}{|V|}\sum_{i \in V} p_i \leq \alpha - \delta \text{ and } \frac{1}{|\hat{V}|}\sum_{i \in \hat{V}} \hat{p}_i \leq \alpha.$$

Then

$$\sum_{i \in \hat{V}} 1 - p_i \geq \frac{1 - \alpha - \delta}{1 - \alpha + \delta + 1/(|V|+1)} \sum_{i \in V} 1 - p_i.$$

Proof. Observe first that $|\hat{V}| \geq |V|$, since

$$\frac{1}{|V|}\sum_{i\in V}\hat{p}_i \leq \frac{1}{|V|}\sum_{i\in V}(p_i+\delta) \leq \alpha.$$

Now observe that

$$\sum_{i\in\hat{V}} 1-p_i \geq \sum_{i\in\hat{V}} 1-\hat{p}_i-\delta = |\hat{V}|(1-\alpha-\delta) \geq |V|(1-\alpha-\delta).$$

On the other hand, by considering adding a single element to V, we have

$$\frac{1}{|V|}\sum_{i\in V} p_i \geq \alpha-\delta-\frac{1}{|V|+1}.$$

Thus,

$$\sum_{i\in V} 1-p_i \leq |V|\left(1-\alpha+\delta+\frac{1}{|V|+1}\right).$$

Taking ratios completes the proof. □

We now turn to the proof of Theorem 5.

Proof of Theorem 5. Let $(x_1,y_1,t_1,h_1),\ldots,(x_n,y_n,t_n,h_n)$ denote the samples from Equation (4) (where $h_1,\ldots,h_n$ are unobserved). From the structure of Equation (4), we have that, conditioned on $(x_1,y_1,t_1),\ldots,(x_n,y_n,t_n)$, the random variables $\mathbb{1}[h_i=0]$ are Bernoulli random variables with biases $p_i = \Pr(h_i=0 \mid x_i,y_i,t_i)$.

From Lemma 13, we have that

$$|\hat{w}_i - \Pr(H=0 \mid x_i,y_i,t_i=1)| \leq \frac{2U(U+2(\lambda+1))}{L^2}\epsilon$$

for all $i=1,\ldots,n$. If S is a set of indices satisfying $t_i=1$ for all $i\in S$ and $\frac{1}{|S|}\sum_{i\in S}\hat{w}_i \leq \alpha$, then

$$\begin{aligned}\mathbb{E}\left[\frac{1}{|S|}\sum_{i\in S}\mathbb{1}[h_i=0] \mid (x_1,y_1,t_1),\ldots,(x_ny_n,t_n)\right] &= \frac{1}{|S|}\sum_{i\in S}\mathbb{E}\left[\mathbb{1}[h_i=0] \mid (x_1,y_1,t_1),\ldots,(x_ny_n,t_n)\right] \\ &= \frac{1}{|S|}\sum_{i\in S}\hat{w}_i + \frac{2U(U+2(\lambda+1))}{L^2}\epsilon \\ &\leq \alpha + \frac{2U(U+2(\lambda+1))}{L^2}\epsilon.\end{aligned}$$

To prove the power statement, observe that the Bayes' optimal procedure is to choose the largest set V such that $\frac{1}{|V|}\sum_{i\in V} p_i \leq \alpha$. Then the power conclusion follows directly from Lemma 14 with $\delta = \frac{2U(U+2(\lambda+1))}{L^2}\epsilon$. □

D Additional results and details from Section 4

D.1 Proof of Lemma 6

Suppose there exist distributions π, f_0, f_1 such that

$$\begin{aligned}&p(y \mid x,t=0) := f_0(y \mid x), \text{ and} \\ &p(y \mid x,t=1) := f_t(y \mid x) = (1-\pi(x))f_0(y \mid x) + \pi(x)f_1(y \mid x).\end{aligned}$$

Pick any x. Observe that if $\pi(x)=0$, then the corresponding $\pi^\star(x)=0$ and the lemma conclusion trivially holds. Thus, we may assume $\pi(x)>0$, in which case we have by rearrangement:

$$f_1(y \mid x) = \frac{1}{\pi(x)}\left(f_t(y \mid x) - f_0(y \mid x)\right) + f_0(y \mid x).$$

As a probability density, $f_1(y \mid x) \geq 0$, which implies

$$\pi(x) \geq 1 - \frac{f_t(y \mid x)}{f_0(y \mid x)}.$$

Since the above holds for all y, we therefore must have

$$\pi(x) \geq 1 - \min_y \frac{f_t(y \mid x)}{f_0(y \mid x)} = \pi^\star(x).$$

To prove the second part of the lemma, let x be a point and $\hat{\pi}$ be a function satisfying $1 \geq \hat{\pi}(x) \geq \pi^\star(x) > 0$. Then define

$$f_1^{\hat{\pi}}(y \mid x) = \frac{1}{\hat{\pi}(x)} \left(f_t(y \mid x) - f_0(y \mid x)\right) + f_0(y \mid x).$$

By definition of $\pi^\star$, we have $f_1^{\hat{\pi}}(y \mid x) \geq 0$. Moreover, integrating the right hand side over y shows that $f_1^{\hat{\pi}}(y \mid x)$ integrates to 1. Thus, $f_1^{\hat{\pi}}(y \mid x)$ is a proper density. Rearrangement shows that

$$f_t(y \mid x) = (1 - \hat{\pi}(x)) f_0(y \mid x) + \hat{\pi}(x) f_1^{\hat{\pi}}(y \mid x).$$

D.2 Proof of Theorem 7

To prove Theorem 7 we first need to prove some lemmas. The first lemma shows that under the conditions of Theorem 7, we can estimate $\pi^\star$ accurately.

Lemma 15. *Let $\epsilon, L, U > 0$ be given. Suppose the following holds for all $i = 1, \ldots n$ and $y \in \mathcal{Y}$:*

(i) $L \leq f_t(y \mid x_i), f_0(y \mid x_i) \leq U$,

(ii) $|f_t(y \mid x_i) - \hat{f}_t(y \mid x_i)| \leq \epsilon$, *and*

(iii) $|f_0(y \mid x_i) - \hat{f}_0(y \mid x_i)| \leq \epsilon$.

If $\epsilon \leq L/2$, then $|\hat{\pi}^\star(x_i) - \pi^\star(x_i)| \leq \frac{8U\epsilon}{L^2}$ for all $i = 1, \ldots, n$, where $\hat{\pi}^\star(x) = 1 - \min_y \frac{\hat{f}_t(y|x)}{\hat{f}_0(y|x)}$.

Proof. Pick some index i, and let $x = x_i$. Let $y = \operatorname{argmin}_y \frac{f_t(y|x)}{f_0(y|x)}$ and $\hat{y} = \operatorname{argmin}_y \frac{\hat{f}_t(y|x)}{\hat{f}_0(y|x)}$. Then we can write

$$\begin{aligned}
\hat{\pi}^\star(x) - \pi^\star(x) &= \frac{f_t(y \mid x)}{f_0(y \mid x)} - \frac{\hat{f}_t(\hat{y} \mid x)}{\hat{f}_0(\hat{y} \mid x)} \\
&= \frac{f_t(y \mid x)}{f_0(y \mid x)} - \frac{f_t(\hat{y} \mid x)}{f_0(\hat{y} \mid x)} + \frac{f_t(\hat{y} \mid x)}{f_0(\hat{y} \mid x)} - \frac{\hat{f}_t(\hat{y} \mid x)}{\hat{f}_0(\hat{y} \mid x)} \\
&\leq \frac{f_t(\hat{y} \mid x)}{f_0(\hat{y} \mid x)} - \frac{\hat{f}_t(\hat{y} \mid x)}{\hat{f}_0(\hat{y} \mid x)} \\
&= \frac{f_t(\hat{y} \mid x)\hat{f}_0(\hat{y} \mid x) - f_0(\hat{y} \mid x)\hat{f}_t(\hat{y} \mid x)}{f_0(\hat{y} \mid x)\hat{f}_0(\hat{y} \mid x)} \\
&\leq \frac{(\hat{f}_t(\hat{y} \mid x) + \epsilon)(f_0(\hat{y} \mid x) + \epsilon) - f_0(\hat{y} \mid x)\hat{f}_t(\hat{y} \mid x)}{f_0(\hat{y} \mid x)\hat{f}_0(\hat{y} \mid x)} \\
&\leq \frac{2\epsilon(U + \epsilon)}{L^2/2} \leq \frac{8U\epsilon}{L^2}.
\end{aligned}$$

The first inequality comes from the definition of y, the second inequality comes from Assumptions (ii) and (iii), and the last two inequalities follow from Assumption (i). A symmetric line of reasoning that demonstrates $\pi^\star(x) - \hat{\pi}^\star(x) \leq \frac{8U\epsilon}{L^2}$ completes the proof. □

Our second lemma shows that the posterior estimate in Equation (6) is close to the ground truth value.

Lemma 16. *Let $\epsilon, L, U > 0$ be given. Suppose the following holds for all $i = 1, \dots n$ and $y \in \mathcal{Y}$:*

(i) $L \le f_t(y \mid x_i), f_0(y \mid x_i) \le U$,

(ii) $|f_t(y \mid x_i) - \hat{f}_t(y \mid x_i)| \le \epsilon$, *and*

(iii) $|f_0(y \mid x_i) - \hat{f}_0(y \mid x_i)| \le \epsilon$.

If $\epsilon \le \min(L/2, 1)$, then for all $i = 1, \dots, n$,

$$\left| \frac{(1 - \pi^\star(x_i)) f_0(y_i \mid x_i)}{f_t(y_i \mid x_i)} - \frac{(1 - \hat{\pi}^\star(x_i)) \hat{f}_0(y_i \mid x_i)}{\hat{f}_t(y_i \mid x_i)} \right| \le \frac{8U}{L^2} \left(1 + \frac{12U}{L^2}\right) \epsilon.$$

Proof. Let $x = x_i$, $y = y_i$. Then we have

$$\begin{aligned}
&\frac{(1 - \pi^\star(x)) f_0(y \mid x)}{f_t(y \mid x)} - \frac{(1 - \hat{\pi}^\star(x)) \hat{f}_0(y \mid x)}{\hat{f}_t(y \mid x)} \\
&\quad = \frac{(1 - \pi^\star(x)) f_0(y \mid x) \hat{f}_t(y \mid x) - (1 - \hat{\pi}^\star(x)) \hat{f}_0(y \mid x) f_t(y \mid x)}{f_t(y \mid x) \hat{f}_t(y \mid x)} \\
&\quad \le \frac{((1 - \hat{\pi}^\star(x)) + \frac{8U\epsilon}{L^2})(\hat{f}_0(y \mid x) + \epsilon)(f_t(y \mid x) + \epsilon) - (1 - \hat{\pi}^\star(x)) \hat{f}_0(y \mid x) f_t(y \mid x)}{f_t(y \mid x) \hat{f}_t(y \mid x)} \\
&\quad \le \frac{(U + \epsilon)\epsilon + U\epsilon + \epsilon^2 + \frac{8U^2\epsilon}{L^2}(U + \epsilon) + \frac{8U\epsilon}{L^2}\epsilon^2(U + \epsilon) + \frac{8U^2\epsilon^2}{L^2} + \frac{8U\epsilon^3}{L^2}}{L^2/2} \\
&\quad \le \frac{8U}{L^2} \left(1 + \frac{12U}{L^2}\right) \epsilon.
\end{aligned}$$

Here, we have used Lemma 15 in the first inequality. □

We now turn to the proof of Theorem 7, which is nearly identical to the proof of Theorem 5.

Proof of Theorem 7. Let $(x_1, y_1, t_1, h_1), \dots, (x_n, y_n, t_n, h_n)$ denote draws from Equation (1) (where $h_1, \dots, h_n$ are unobserved). From the structure of Equation (1), we have that, conditioned on $(x_1, y_1, t_1), \dots, (x_n, y_n, t_n)$, the random variables $\mathbb{1}[h_i = 0]$ are Bernoulli random variables with biases $p_i = \Pr(h_i = 0 \mid x_i, y_i, t_i)$.

From Lemma 16, we have that

$$\hat{w}_i \le \Pr(H = 0 \mid x_i, y_i, t_i = 1) + \frac{8U}{L^2} \left(1 + \frac{12U}{L^2}\right) \epsilon$$

for all $i = 1, \dots, n$. If S is a set of indices satisfying $t_i = 1$ for all $i \in S$ and $\frac{1}{|S|} \sum_{i \in S} \hat{w}_i \le \alpha$, then

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{|S|} \sum_{i \in S} \mathbb{1}[h_i = 0] \mid (x_1, y_1, t_1), \dots, (x_n y_n, t_n)\right] &= \frac{1}{|S|} \sum_{i \in S} \mathbb{E}\left[\mathbb{1}[h_i = 0] \mid (x_1, y_1, t_1), \dots, (x_n y_n, t_n)\right] \\
&= \frac{1}{|S|} \sum_{i \in S} \hat{w}_i + \frac{8U}{L^2} \left(1 + \frac{12U}{L^2}\right) \epsilon \\
&\le \alpha + \frac{8U}{L^2} \left(1 + \frac{12U}{L^2}\right) \epsilon.
\end{aligned}$$

To prove the power statement, we first observe that the Bayes' optimal procedure is to select the largest set such that $\frac{1}{|V|} \sum_{i \in V} p_i \le \alpha$, where $p_i = \Pr(h_i = 0 \mid x_i, y_i, t_i)$ is formed using the prior $\pi^\star$ from Lemma 6. Then Lemma 14 finishes the argument. □

E Additional results and details from Section 5

E.1 Proof of Proposition 9

We will prove Proposition 9 by breaking it into the following five components.

Proposition 17. *The nonparametric C2G model is well-specified whenever*

$$Y \perp\!\!\!\perp T|X, H{=}0.$$

Proof. Observe that if $Y \perp\!\!\!\perp T|X, H{=}0$, then we have

$$p(y \mid x, t{=}0) = p(y \mid x, t{=}0, h{=}0) = p(y \mid x, t{=}1, h{=}0),$$

where the first equality follows from the C2G model and the second follows from $Y \perp\!\!\!\perp T|X, H{=}0$.

Now observe

$$\begin{aligned} p(y \mid X = x, T{=}1) &= \sum_{h=0}^{1} p(H = h \mid X = x, T{=}1) p(y \mid X = x, T{=}1, H = h) \\ &= (1 - \pi(x)) f_0(y \mid x) + \pi(x) f_1(y \mid x). \end{aligned}$$

Here, we have defined

$$\begin{aligned} \pi(x) &= p(H{=}1 \mid X = x, T{=}1) \\ f_0(y \mid x) &= p(Y = y \mid X = x, T{=}0) = p(Y = y \mid X = x, T{=}1, H{=}0) \\ f_1(y \mid x) &= p(Y = y \mid X = x, H{=}1, T{=}1). \end{aligned}$$

□

Proposition 18. *The nonparametric C2G model is not well-specified in general when the scenario does not imply*

$$Y \perp\!\!\!\perp T|X, H{=}0.$$

Proof. By assumption, there is a distribution consistent with the scenario such that Y and T are dependent given $X, H{=}0$. In this case, there is a set of x with positive probability such that

$$p(y|x, T{=}0, H{=}0) \neq p(y|x, T{=}1, H{=}0).$$

Since under the C2G model, we have that $T = 0$ implies $H{=}0$, this implies

$$p(y|x, T{=}0) = p(y|x, T{=}0, H{=}0) \neq p(y|x, T{=}1, H{=}0),$$

and so the nonparametric model is not well-specified. □

Proposition 19. *The additive C2G model is well-specified whenever*

$$Y \perp\!\!\!\perp T|X, H{=}0$$

and there exist functions g, μ_0, μ_1 such that

$$p(y|T{=}1, H{=}1, x) = g(y - \mu_1(x)) = p(y + \mu_0(x) - \mu_1(x)|T{=}0, x)$$

for every x.

Proof. The proof follows directly from the fact that Proposition 17 implies that Equation (7) holds and the condition

$$p(y|T{=}1, H{=}1, x) = g(y - \mu_1(x)) = p(y + \mu_0(x) - \mu_1(x)|T{=}0, x)$$

implies that Equation (4) holds. □

Proposition 20. *The additive C2G model is not well-specified in general when the scenario does not imply both $Y \perp\!\!\!\perp T \mid X, H{=}0$ and the existence of functions g, μ_0, μ_1 such that*

$$p(y \mid T{=}1, H{=}1, x) = g(y - \mu_1(x)) = p(y + \mu_0(x) - \mu_1(x) \mid T{=}0, x)$$

for every x.

Proof. Observe that if the scenario does not imply $Y \perp\!\!\!\perp T \mid X, H{=}0$, then by Proposition 18, Equation (7) does not hold and so the additive C2G model is not well-specified.

On the other hand, if the scenario does not imply

$$p(y \mid T{=}1, H{=}1, x) = g(y - \mu_1(x)) = p(y + \mu_0(x) - \mu_1(x) \mid T{=}0, x)$$

then Equation (4) does not hold on a set of x with positive probability. □

Proposition 21. *Suppose $Y \perp\!\!\!\perp T|X, U, H$. If either (a) $Y \perp\!\!\!\perp U \mid X, H{=}0$ or (b) $U \perp\!\!\!\perp (T, H) \mid X$, then $Y \perp\!\!\!\perp T \mid X, H{=}0$.*

Proof. Observe that under $Y \perp\!\!\!\perp T|X, U, H$, we have

$$p(y \mid x, t, H{=}0) = \int p(y \mid x, u, H{=}0) p(u \mid x, t, H{=}0)\, du.$$

If (a) holds, then $p(y \mid x, u, H{=}0) = p(y \mid x, H{=}0)$ and so

$$p(y \mid x, t, H{=}0) = p(y \mid x, H{=}0) \int p(u \mid x, t, H{=}0)\, du = p(y \mid x, H{=}0),$$

and so $Y \perp\!\!\!\perp T \mid X, H{=}0$.

On the other hand, if (b) holds, then $p(u \mid x, t, H{=}0) = p(u|x)$ and so

$$p(y \mid x, t, H{=}0) = \int p(y \mid x, u, H{=}0) p(u|x)\, du = f(x, y)$$

for some function f that does not depend on t, which implies $Y \perp\!\!\!\perp T \mid X, H{=}0$. □

E.2 Proof of Corollary 10

To prove Corollary 10, we break it into the following results.

Proposition 22. *Nonparametric C2G is well-specified under response confounding. Additionally, if $p(y|x, h)$ satisfies Equation* (4)*, additive C2G is also well-specified.*

Proof. Observe that under response confounding, Y's only parents are X and H, which gives us $Y \perp\!\!\!\perp T|X, H$. In particular this implies $Y \perp\!\!\!\perp T|X, H{=}0$. Proposition 9 finishes the proof. □

Proposition 23. *Nonparametric C2G is well-specified under effect-size confounding.*

Proof. Observe that under effect-size confounding, whenever $H = 0$, we have $Y = Y_0$. Moreover, Y_0's only parent is X, which gives us $Y_0 \perp\!\!\!\perp (T, H) \mid X$, and therefore that $Y_0 \perp\!\!\!\perp T \mid X, H{=}0$. Thus, we have $Y \perp\!\!\!\perp T \mid X, H{=}0$. Proposition 9 finishes the proof. □

Proposition 24. *Under effect-size confounding, the additive C2G model is not well-specified in general, even when $Y_0 = \mu_0(X) + \epsilon$ with $\epsilon \sim g$.*

Proof. Say $\epsilon \sim \mathcal{N}(0, 1)$ and $U \sim \text{Bern}(1/2)$. Let

$$H \mid X{=}x, U{=}u, T{=}1 \sim \text{Bern}(1/2).$$

Further, let us take $Y_0 = \epsilon$ and $\Delta = U$. Observe that by Bayes' rule,

$$\Pr(U{=}1 \mid x, T{=}1, H{=}1) = 1/2.$$

Thus we have

$$Y \mid x, T{=}0 \sim \mathcal{N}(0, 1),$$
$$Y \mid x, T{=}1, H{=}1 \sim \tfrac{1}{2}\mathcal{N}(0, 1) + \tfrac{1}{2}\mathcal{N}(1, 1).$$

The variance of $Y \mid x, T{=}0$ is 1 but the variance of $Y \mid x, T{=}1, H{=}1$ is 5/4. Since a location-shift preserves variance, Equation (4) is not satisfied, and the additive C2G is not well-specified here. □

Proposition 25. *Under total effect confounding, neither the nonparametric nor the additive C2G model is well-specified in general.*

Proof. Let X be constant, and consider the instance of total effect confounding given by

$$\begin{aligned} U &\sim \text{Bern}(1/2) \\ T &\sim \text{Bern}(1/2) \text{ independent of } U \\ H \mid U{=}u, T{=}1 &\sim \text{Bern}(\pi_u) \\ Y \mid U{=}u, H{=}0 &\sim \mathcal{N}(u, 1) \end{aligned}$$

with $\pi_0 = 1/4$ and $\pi_1 = 3/4$. Since $T \perp\!\!\!\perp U$, $\Pr(U{=}1 \mid T{=}0) = 1/2$. By Bayes' rule,

$$\Pr(U{=}1 \mid T{=}1, H{=}0) = \frac{\frac{1}{2}(1-\pi_1)}{\frac{1}{2}(1-\pi_0) + \frac{1}{2}(1-\pi_1)} = \frac{1}{4}.$$

Thus,

$$\begin{aligned} \mathbb{E}[Y \mid T{=}0] &= \mathbb{E}[Y \mid H{=}0, T{=}0] \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid H{=}0, T{=}0, U] \mid H{=}0, T{=}0\right] \\ &= \mathbb{E}\left[U \mid H{=}0, T{=}0\right] = 1/2. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbb{E}[Y \mid T{=}1, H{=}0] &= \mathbb{E}\left[\mathbb{E}[Y \mid H{=}0, T{=}1, U] \mid H{=}0, T{=}1\right] \\ &= \mathbb{E}\left[U \mid H{=}0, T{=}1\right] = 1/4. \end{aligned}$$

Hence $Y \not\perp\!\!\!\perp T \mid H{=}0$. The statement then follows from Propositions 18 and 20. □

Proposition 26. *Under canonical confounding, neither the nonparametric nor the additive C2G model is well-specified in general.*

Proof. Let X be constant, and consider the instance of canonical confounding given by

$$\begin{aligned} U &\sim \text{Bern}(1/2) \\ T \mid U{=}u &\sim \text{Bern}(\pi_u) \\ H \mid T{=}1 &\sim \text{Bern}(1/2) \text{ independent of } U \\ Y \mid U{=}u, H{=}0 &\sim \mathcal{N}(u, 1) \end{aligned}$$

with $\pi_0 = 1/4$ and $\pi_1 = 3/4$. By Bayes' rule,

$$\Pr(U{=}1 \mid T{=}0) = \frac{\frac{1}{2}(1-\pi_1)}{\frac{1}{2}(1-\pi_0) + \frac{1}{2}(1-\pi_1)} = \frac{1}{4}.$$

Since $H \perp\!\!\!\perp U \mid T{=}1$, Bayes' rule also gives

$$\Pr(U{=}1 \mid T{=}1, H{=}0) = \Pr(U{=}1 \mid T{=}1) = \frac{\frac{1}{2}\pi_1}{\frac{1}{2}\pi_0 + \frac{1}{2}\pi_1} = \frac{3}{4}.$$

Since Y's parents are U and H, we have $\mathbb{E}[Y \mid H{=}0, T{=}t, U] = U$ for $t \in \{0, 1\}$. Thus,

$$\begin{aligned} \mathbb{E}[Y \mid T{=}0] &= \mathbb{E}[Y \mid H{=}0, T{=}0] \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid H{=}0, T{=}0, U] \mid H{=}0, T{=}0\right] \\ &= \mathbb{E}\left[U \mid H{=}0, T{=}0\right] = 1/4. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbb{E}[Y \mid T{=}1, H{=}0] &= \mathbb{E}\left[\mathbb{E}[Y \mid H{=}0, T{=}1, U] \mid H{=}0, T{=}1\right] \\ &= \mathbb{E}\left[U \mid H{=}0, T{=}1\right] = 3/4. \end{aligned}$$

Hence $Y \not\perp\!\!\!\perp T \mid H{=}0$. The statement then follows from Propositions 18 and 20. □

E.3 A stronger counterexample for canonical confounding

We will make use of the following lemma.

Lemma 27. *Let $\mathcal{D}$ and $\mathcal{D}'$ be two distributions. For random variables X, Y, H, T, U, there is a choice of distributions for U, $Y|X,H,U$, $T|X,U$, and $H|X,T$ such that*

1. *X, Y, H, T, U obey the canonical confounding graphical model,*
2. *$\Pr(T{=}1 \mid X) = 1/2$ [no treatment overlap violations],*
3. *$\Pr(H{=}0 \mid X, T{=}0) = 1$ [all untreated are non-responders],*
4. *$\Pr(H{=}1 \mid X, T{=}1) = 1/2$ [no responder overlap violations],*
5. *$Y \mid X, T{=}0 \sim \mathcal{D}'$, and*
6. *$Y \mid X, H{=}0, T{=}1 \stackrel{d}{=} Y \mid X, H{=}1, T{=}1 \stackrel{d}{=} Y \mid X, T{=}1 \sim \mathcal{D}$ [matching responder and non-responder distributions].*

Proof. Our conditional distributions will be independent of X, so we will drop X going forward. We make the following choices for the conditional distributions

$$\begin{aligned} Y \mid H{=}0, U = 0 &\sim \mathcal{D} \\ Y \mid H{=}0, U = 1 &\sim \mathcal{D}' \\ Y \mid H{=}1, U = 0 &\sim \mathcal{D} \\ Y \mid H{=}1, U = 1 &\sim \mathcal{D}' \\ \Pr(T{=}1 \mid U = 0) &= 1 \\ \Pr(T{=}1 \mid U = 1) &= 0 \\ \Pr(H{=}1 \mid T{=}1) &= 1/2 \\ \Pr(H{=}0 \mid T{=}0) &= 1 \\ \Pr(U = 1) &= 1/2. \end{aligned}$$

Then we can work out the following

$$\begin{aligned} \Pr(T{=}1) &= \Pr(U = 1)\Pr(T = 1 \mid U = 1) + \Pr(U = 0)\Pr(T = 1 \mid U = 0) = 1/2 \\ \Pr(U = 1 \mid H, T{=}1) &= \Pr(U = 1 \mid T{=}1) \\ &= \frac{\Pr(T{=}1 \mid U = 1)\Pr(U = 1)}{\Pr(T{=}1 \mid U = 1)\Pr(U = 1) + \Pr(T{=}1 \mid U = 0)\Pr(U = 0)} = 0 \\ \Pr(U = 1 \mid H, T{=}0) &= \Pr(U = 1 \mid T{=}0) \\ &= \frac{\Pr(T{=}0 \mid U = 1)\Pr(U = 1)}{\Pr(T{=}0 \mid U = 1)\Pr(U = 1) + \Pr(T{=}0 \mid U = 0)\Pr(U = 0)} = 1 \\ \Pr(U = 0 \mid H, T{=}1) &= 1 \\ \Pr(U = 0 \mid H, T{=}0) &= 0. \end{aligned}$$

And so we have

$$\begin{aligned} \Pr(Y = y|H{=}1, T{=}1) &= \Pr(Y = y \mid H{=}1, U = 0, T{=}1)\Pr(U = 0 \mid H{=}1, T{=}1) \\ &\quad + \Pr(Y = y \mid H{=}1, U = 1, T{=}1)\Pr(U = 1 \mid H{=}1, T{=}1) \\ &= \Pr(Y = y \mid H{=}1, U = 0)\Pr(U = 0 \mid H{=}1, T{=}1) \\ &\quad + \Pr(Y = y \mid H{=}1, U = 1)\Pr(U = 1 \mid H{=}1, T{=}1) \\ &= \Pr(Y = y \mid H{=}1, U = 0). \end{aligned}$$

Thus, $Y \mid H{=}1, T{=}1 \sim \mathcal{D}$. Similarly, we have

$$\begin{aligned} \Pr(Y = y|H{=}0, T{=}1) &= \Pr(Y = y \mid H{=}0, U = 0, T{=}1)\Pr(U = 0 \mid H{=}0, T{=}1) \\ &\quad + \Pr(Y = y \mid H{=}0, U = 1, T{=}1)\Pr(U = 1 \mid H{=}0, T{=}1) \\ &= \Pr(Y = y \mid H{=}0, U = 0)\Pr(U = 0 \mid H{=}0, T{=}1) \\ &\quad + \Pr(Y = y \mid H{=}0, U = 1)\Pr(U = 1 \mid H{=}0, T{=}1) \\ &= \Pr(Y = y \mid H{=}0, U = 0). \end{aligned}$$

Therefore, $Y \mid H{=}0, T{=}1 \sim \mathcal{D}$. Moreover, we can work out

$$\begin{aligned}\Pr(Y = y \mid T{=}0) &= \Pr(Y = y \mid H{=}0, T{=}0) \\ &= \Pr(Y = y \mid U = 0, H{=}0)\Pr(U = 0 \mid T{=}0) \\ &\quad + \Pr(Y = y \mid U = 1, H{=}0)\Pr(U = 1 \mid T{=}0) \\ &= \Pr(Y = y \mid U = 1, H{=}0).\end{aligned}$$

Thus, $Y \mid T{=}0 \sim \mathcal{D}'$. □

To see how this result implies that the C2G models are not well-specified under canonical confounding, we can take $\mathcal{D}$ and $\mathcal{D}'$ to be distributions with non-intersecting supports, e.g. uniform distributions over separated intervals. Then Lemma 27 gives us a confounded setting in which the non-treatment distribution does not match the non-response-under-treatment distribution, violating Equation (7).

Moreover, we can also see that if we were to apply the nonparametric causal two-group model anyway, the procedure would estimate $\pi^\star(x) = 1$ for all x. This is because the treatment and non-treatment distributions do not share any support, and so the only valid mixture satisfying Equation (2) is the trivial mixture that assigns the responder distribution to be the treatment distribution itself and for $H = 1$ to occur with probability 1 when $T = 1$. However, from Lemma 27, the true probability of $H = 1$ is 1/2 when $T = 1$.

F Additional results and details from Section 6

Additive	Nonadditive
$\beta, \gamma, \theta \sim \mathcal{N}(0, I_d/\sqrt{d})$	$\beta, \gamma \sim \mathcal{N}(0, I_d/\sqrt{d})$
	$c \sim \text{HalfNormal}(0, 2)$
$H\|T = 1 \sim \text{Bern}(\text{sigmoid}(\beta^T X))$	$W = \gamma^T X$
$T \sim \text{Bern}(0.5)$	$T \sim \text{Bern}(\text{sigmoid}(W)$
$\mu_0 = \gamma^T X$	$H\|T = 1 \sim \text{Bern}(\text{sigmoid}(\beta^T X))$
$\mu_1 = \mu_0 + \tau \sum_i \lvert X_i \rvert \lvert \gamma_i \rvert$	$B_{ij} \sim \text{Bern}(0.1), Z_{ij} \sim \text{StudentT}(3)$
$Y\|H = 0 \sim \mathcal{N}(\mu_0, 1)$	$L = \sum_{i,j} B_{i,j} Z_{i,j} X_i X_j$
$Y\|H = 1 \sim \mathcal{N}(\mu_1, 1)$	$Y \sim \mathcal{N}(\log(1 + \exp(cW + \theta^T X + \tau H + L)), 1)$

Table 4: Data generating models for additive and nonadditive synthetic simulations.

The full data generating process for both synthetic additive and nonadditive simulations is described in Table 4. All simulations were performed on a cluster with Intel Xeon Gold 6348 CPUs. Each simulation ran in under 2 hours with 2 CPUs allocated. In all plots, shaded regions represent 95% confidence intervals under the normal approximation.

For additional baselines, we also compared against three methods commonly used to test for treatment effects. FDR-Regression [45] controls the local FDR in different subsets of the experiment. BART [18] and Causal forests [53] are two nonparametric tree-based methods for estimating individual treatment effects.

To appropriately measure power in this setting, we considered *valid power*, which we define to be

$$\text{vp}(\alpha) = \begin{cases} 0 & \text{if } \hat{\text{fdr}}(\alpha) - \text{CI}_{\text{fdr}} > \alpha \\ \frac{\#\text{ of rejected true non-nulls}}{\#\text{of true non-nulls}} & \text{otherwise} \end{cases},$$

where $\hat{\text{fdr}}(\alpha)$ is the average observed FDR of the procedure at level α and CI_{fdr} is its associated 95% confidence interval band, over the 50 different random runs. Intuitively, valid power captures the standard notion of power in settings where FDR has been respected.

We generally observe pathological behavior for these baselines. In particular, both BART and Causal forest generally reject everything at extremely low nominal FDR levels, leading to high observed FDR rates. The valid power for these methods is categorically zero until the first nominal FDR level for which the observed FDR is acceptable, at which point the valid power immediately jumps to 1. On the other hand, we observe that FDRreg's behavior changes between settings, having high FDR

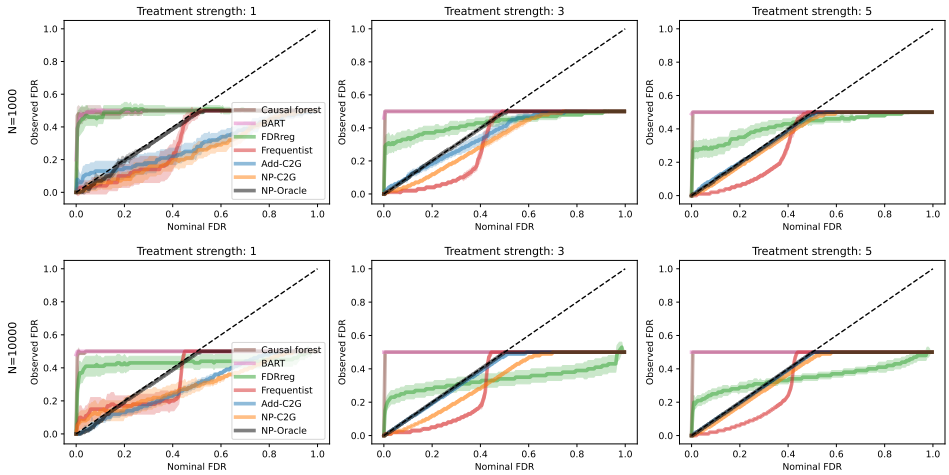


Figure 3: FDR curves on additive synthetic data.

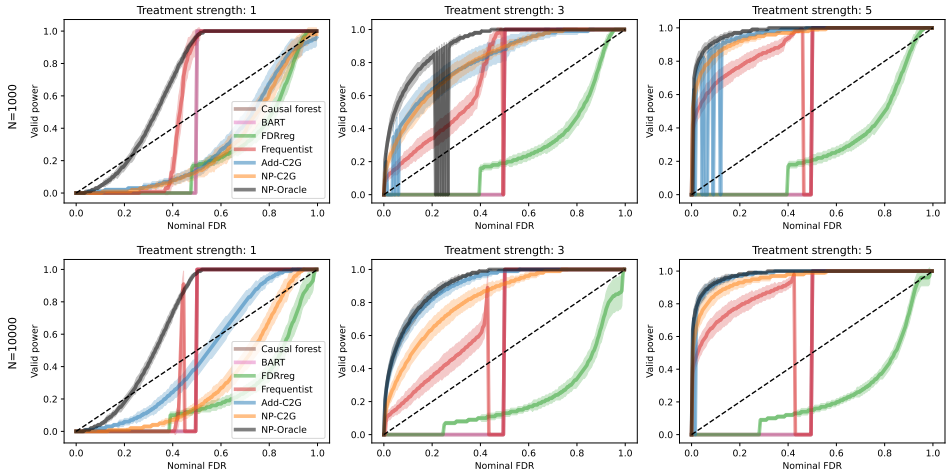


Figure 4: Valid power curves on additive synthetic data.

values on the additive data and low FDR values on the nonadditive data. In all settings, however, the valid power of FDRreg was far below the C2G methods and the frequentist baseline.

It is worth emphasizing that these results are not intended as an indictment of the baseline methods. When the assumptions of the problem match those of the methods, these baselines generally perform quite well. Rather, these results illustrate how the causal two-groups setting breaks the assumptions behind these methods.

G Semi-synthetic simulations on drug sensitivity data

The Genomics of Drug Sensitivity in Cancer (GDSC) is a repository of cancer cell line data, both genomic and drug-response [21, 58]. We considered a subset of 832 cell lines for which both genomic data and dose-response data against the drug Nutlin-3a were recorded. Nutlin-3a is an MDM2-inhibitor, and as such it promotes p53, a tumor suppressing protein [47]. However, this pathway can successfully inhibit cancer growth only when the corresponding gene, TP53, is wild-type, i.e. does not contain a missense mutation.

For each cell line in our dataset, we took the covariates X to be the 50-dimensional PCA projection of its RNA expression data. We took the outcome variable Y to be the z-scored cell line viability of the cell line when exposed to the maximum dose of Nutlin-3a. We let H be an indicator variable that

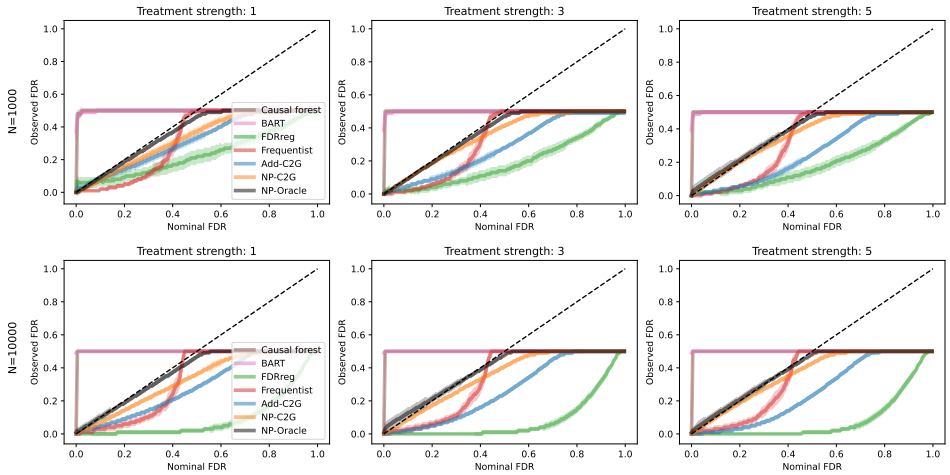


Figure 5: FDR curves on nonadditive synthetic data.

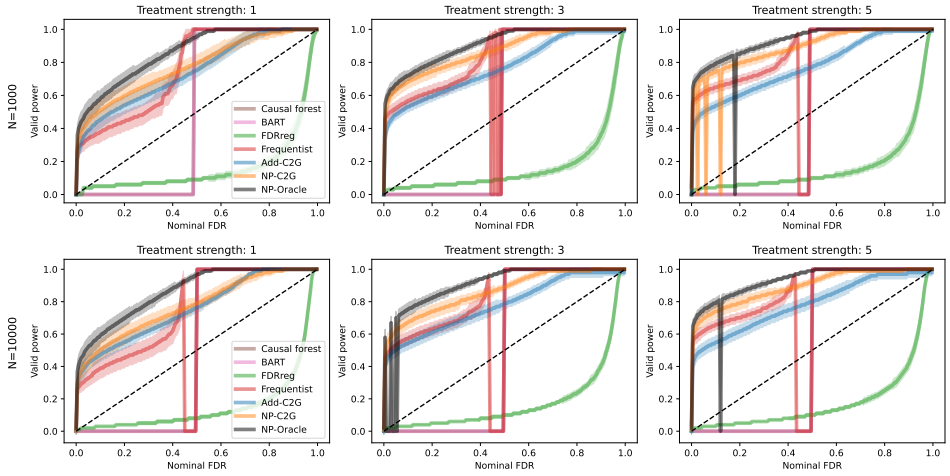


Figure 6: Valid power curves on nonadditive synthetic data.

is 1 when the cell line does not have a TP53 mutation. And we took T to be 1 whenever $H = 1$, and otherwise we set it to a Bernoulli draw with bias equal to sigmoid(W_{GAPDH}), where W_{GAPDH} is the RNA expression of the Glyceraldehyde 3-phosphate dehydrogenase gene. We reran our simulations with 50 different random seeds.

Table 6 displays the empirical results on GDSC data. Note that in this setting, all methods did a reasonable job of controlling FDR, save for Add-C2G at the lower levels of α. However, for BART, FDRreg, and Causal Forests, this is mostly due to having an unreasonably low rejection rate. The fact that Add-C2G generally does not control FDR in this setting likely indicates that the additivity assumption is grossly violated here. Finally, we see that NP-C2G controlled FDR at all levels while maintaing high valid power.

H Additional results and details from Section 7

Each patient in the immunotherapy dataset is associated with a binary vector where one indicates a somatic mutation in a particular gene and zero indicates that gene is not mutated. Because the gene panels changed over the course of the study, there are missing entries in the patient-by-gene matrix. To handle this, we performed Bernoulli matrix factorization on the sequencing matrix to create 5-dimensional embeddings for each patient, where the dimension was chosen using 3-fold

	Additive			Nonadditive		
N	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 1$	$\tau = 3$	$\tau = 5$
1K	0.19±0.03	0.63±0.05	0.81±0.03	0.67±0.05	0.66±0.06	0.63±0.06
10K	0.11±0.01	0.6±0.04	0.85±0.03	0.66±0.05	0.75±0.06	0.69±0.07

Table 5: Average response effect (ARE) interval results on synthetic data for NP-C2G. Values indicate Jaccard index with the oracle ARE interval; ± denotes 95% confidence intervals.

FDR			
Method	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.25$
Frequentist	**0.03±0.006**	**0.04±0.007**	**0.19±0.038**
Add-C2G	0.14±0.035	0.17±0.034	**0.23±0.032**
NP-C2G	**0.04±0.009**	**0.07±0.011**	**0.23±0.02**
Valid power			
Method	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.25$
Frequentist	0.24±0.018	0.33±0.021	**0.63±0.062**
Add-C2G	0.0±0.0	0.0±0.0	0.42±0.06
NP-C2G	**0.33±0.025**	**0.44±0.023**	**0.65±0.021**

Table 6: Empirical results on semi-synthetic drug-response data. ± denotes 95% confindence intervals. **Bolded** results on FDR column indicate that the method(s) achieved valid FDR (up to 95% confidence intervals). **Bolded** results on valid power column indicate that the method(s) achieved the highest valid power for that setting (up to 95% confidence intervals).

cross-validation on held out log-likelihood. The embeddings, combined with the clinical features of sex, age, cancer type and metastatic status, form the confounders in our analysis.

To transform the survival information to numerical outcomes, we first fit a Cox proportional hazards model to the monotherapy (untreated) group. For each patient i, we let s_i denote the patient's survival time if they were uncensored. If they were censored, s_i is the expectation of this time under the Cox model conditioned on being at least the observed survival time. This provides a conservative estimate of the true survival time because the conditional expectation will shrink survival times to the mean. We convert survival times to z-scores via the inverse normal cumulative distribution function $P(\text{survival time} \leq s_i)$ under the Cox model. We use the z-score as the outcome variable y_i for each patient. Figure 8 shows the survival curves for both groups and the resulting outcome variables.

Figure 9 shows the q-values of individual genes as a function of the nominal FDR level used in NP-C2G selection.

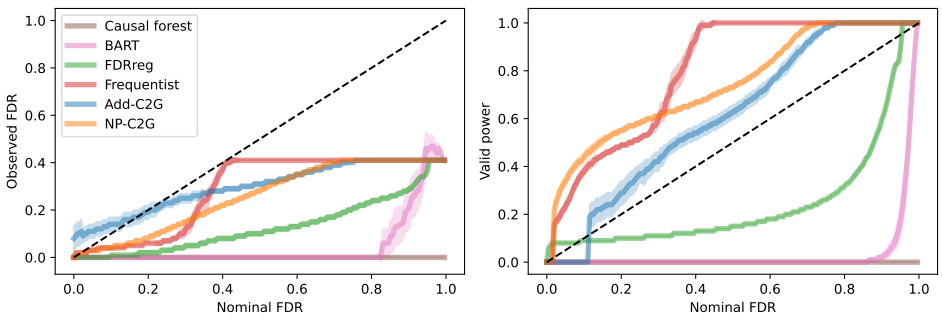


Figure 7: Observed FDR (left) and valid power curves (right) for semi-synthetic drug-response data.

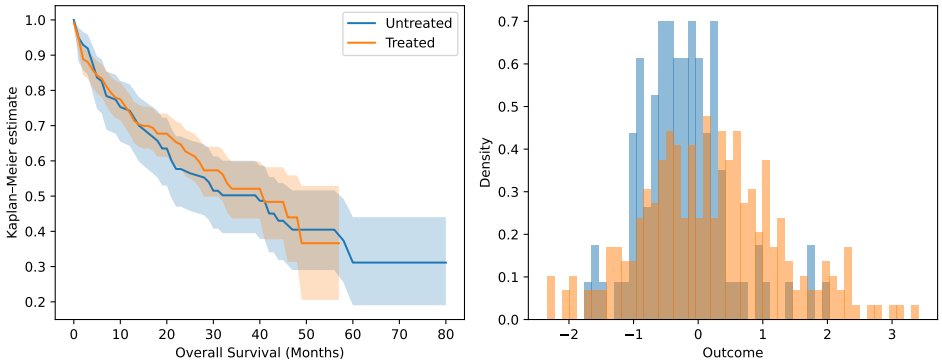


Figure 8: Case study survival data. *Left:* Kaplan-Meier fits to the survival data from the untreated and treated groups. *Right:* Histogram of Cox proportional hazard-transformed outcomes for the untreated and treated groups.

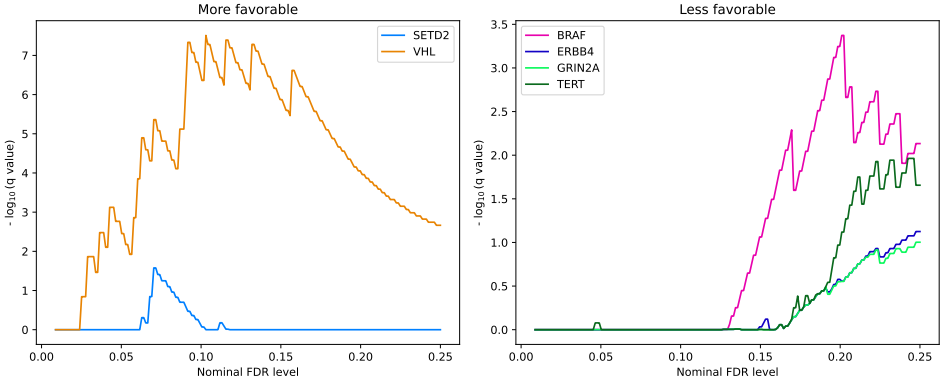


Figure 9: q-values as a function of nominal FDR level for each gene. Only genes achieving q-value below 0.1 are shown.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction outline the contributions of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: All assumptions of our results are clearly stated, and we have a separate limitations paragraph emphasizing that our algorithmic guarantees may fail if those assumptions do not hold.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are in the theorem statements. All proofs are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental conditions are provided in the paper/appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and data are provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training details are provided in the paper/appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error bars/confidence intervals are suitably explained.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The amount of compute needed is outlined in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper abides by the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper does not introduce any concepts that the authors foresee as introducing negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There are no such risks in this paper.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: n/a

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing/human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No crowdsourcing/human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This paper does not use LLMs as a component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.