Controllable Generative Sandbox for Causal Inference

Qi Zhang
Emory University
qi.zhang2@emory.edu

Harsh Parikh
Yale University
harsh.parikh@yale.edu

Ashley I. Naimi
Emory University
ashley.naimi@emory.edu

Razieh Nabi
Emory University
razieh.nabi@emory.edu

Christopher Kim
Amgen
chrkim@amgen.com

Timothy L. Lash
Emory University
timothy.lee.lash@emory.edu

Abstract

Method validation and study design in causal inference rely on synthetic data with known counterfactuals. Existing simulators trade off *distributional realism*, the ability to capture mixed-type and multimodal tabular data, against *causal controllability*, including explicit control over overlap, unmeasured confounding, and treatment effect heterogeneity. We introduce CAUSALMIX, a variational generative framework that closes this gap by coupling a mixture of Gaussian latent priors with data-type-specific decoders for continuous, binary, and categorical variables. The model incorporates explicit causal controls: an overlap regularizer shaping propensity-score distributions, alongside direct parameterizations of confounding strength and effect heterogeneity. This unified objective preserves fidelity to the observed data while enabling factorial manipulation of causal mechanisms, allowing overlap, confounding strength, and treatment effect heterogeneity to be varied independently at design time. Across evaluation settings, CAUSALMIX achieves strong distributional fidelity on mixed-type tables while providing stable, fine-grained causal control. We demonstrate practical utility in a comparative safety study of metastatic castration-resistant prostate cancer treatments, using CAUSALMIX to benchmark estimators under calibrated data-generating processes, with additional applications to hyperparameter tuning and power analysis under targeted treatment effect heterogeneity.

1 Introduction

For each unit, only one potential outcome is observed, rendering individual causal effects fundamentally unobservable [1]. Consequently, assessing causal methods requires settings in which the ground truth is known. Synthetic data therefore serves as a workhorse for causal methodology: it supplies counterfactual ground truth, standardized benchmarks, and controllable stress-tests of identification assumptions. In such *sandbox* settings, researchers can probe robustness to confounding, model misspecification and limited overlap, and tune hyperparameters under conditions that mirror intended applications [2, 3]. Real data-inspired simulations also support prospective study design (e.g., power analysis) and privacy-preserving collaboration by releasing realistic, non-identifiable surrogates [4].

However, performance on synthetic data often fails to translate to real-world settings because estimator behavior depends critically on how well modeling assumptions align with properties of the data, including dependence structure, nonlinearity, confounding strength, overlap, and outcome variability. An ideal benchmark must *both* provide access to causal ground truth *and* faithfully reproduce key features of the real data-generating process (DGP) [5, 6].

Gaps. Despite significant advances in deep generative architectures for causal data [7, 8, 9], existing approaches continue to face a fundamental challenge: jointly achieving realistic modeling of mixed-

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

type observational data and explicit, design-time control over key causal properties such as effect heterogeneity, confounding, and overlap. As a result, gains in empirical realism often come at the expense of causal transparency, limiting their usefulness for method evaluation and study planning.

Contributions. We introduce CAUSALMIX, a generative framework that learns a data-generating process calibrated to observed data while exposing explicit causal control parameters. The goal is not to replace domain-specific simulators, but to provide a *calibrated causal sandbox* in which researchers can systematically vary causal mechanisms while preserving realistic data structure.

- **Mixed-type fidelity.** A hybrid latent architecture with a Bayesian Gaussian–mixture prior and data-type-specific decoder heads captures multimodal dependencies in observational tabular data.
- **Targeted causal control.** We expose *design-time* controls for heterogeneous treatment effects, propensity overlap, and unmeasured confounding, enabling targeted manipulation of causal mechanisms.
- **Unified objective.** A joint training objective balances distributional fidelity with causal constraints, allowing realistic data generation without sacrificing causal controllability.
- **Evaluation and applications.** We develop a unified evaluation framework and demonstrate how calibrated generators support estimator benchmarking, with additional applications to hyperparameter tuning and study design.

Real-world observational data provide realism but lack ground-truth counterfactuals, while purely synthetic simulators provide ground truth but often fail to capture complex data structure. Data-fit generators offer a complementary paradigm: they enable evaluation under realistic distributions while retaining access to controlled causal ground truth. CAUSALMIX is designed for this setting, providing a practical tool for benchmarking and study design in causal inference.*

2 Background

Setup. We observe i.i.d. samples $O = (X, T, Y)$, where X denotes covariates, $T \in \{0, 1\}$ a binary treatment, and Y the observed outcome. Let $Y(1)$ and $Y(0)$ be the potential outcomes, with $Y = T\,Y(1) + (1 - T)\,Y(0)$ by consistency. We define the conditional outcome regression $\mu_t(x) \triangleq \mathbb{E}[Y(t) \mid X = x]$ and the propensity score $e(x) \triangleq \mathbb{P}(T = 1 \mid X = x)$ [10], and consider the average and conditional average treatment effects $\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$ and $\text{CATE}(x) = \mu_1(x) - \mu_0(x)$.

Identification of ATE and CATE from observational data typically relies on consistency, unconfoundedness (ignorability) given X, and positivity (overlap), $0 < e(x) < 1$. Unmeasured confounding violates ignorability, and limited overlap forces extrapolation of counterfactual outcomes. Both undermine point identification, inflate variance, and make the finite-sample behavior of estimators depend strongly on confounding strength, model complexity, and overlap, which complicates method comparison.

Synthetic data for causal inference. Existing approaches differ in how tightly they are grounded in real data, whether they support mixed-type variables, and how directly users can control causal properties (summarized in Appendix A). (i) *Parametric simulators* specify structural equations for (X, T, Y), allowing full control over functional forms and noise, but can produce unrealistic data and method rankings that depend heavily on modeling assumptions, especially for mixed-type and multimodal tabular data. (ii) *Semi-synthetic benchmarks* reuse real covariates while simulating $T \mid X$ and $Y \mid T, X$ from fixed mechanisms, offering reproducibility but limited control over causal properties, with potential privacy constraints [11, 2, 12, 13]. (iii) *Data-fit generators* learn distributions from real data, improving realism but often lacking explicit and reliable control over causal mechanisms [14, 15, 16, 17, 18, 19, 20]. Neural approaches such as REALCAUSE and WGAN-based generators emphasize distributional fidelity but offer no explicit, design-time control over unmeasured confounding or user-specified treatment-effect functions, and are therefore not directly comparable in controlled benchmarking [7, 8]. CREDENCE [9] is the closest prior framework: it anchors generation to an observed sample and allows user-specified treatment effects and confounding bias, but it does not control overlap explicitly, is limited for realistic mixed-type tabular data, and does not systematically validate the realized causal structure.

*Code available at `https://github.com/zhangqiecho/causalmix`

Desiderata for benchmarks. Effective benchmarks should therefore capture realistic mixed-type structure, expose effect heterogeneity, confounding, and overlap as explicit design parameters, and jointly optimize distributional fidelity and causal control. CAUSALMIX is designed to meet all three.

3 Methods

3.1 Causal-control problem formulation

We observe samples $O = (X, T, Y)$ from an empirical observational distribution, where X denotes pre-treatment covariates, $T \in \{0, 1\}$ is treatment, and Y is the observed outcome. Our goal is to learn a generator G_θ that produces synthetic samples $O' = (X', T', Y'(0), Y'(1), Y')$ whose observed distribution is close to the empirical distribution, while exposing user-controlled causal structure. The observed outcome is assembled by consistency, $Y' = T'Y'(1) + (1 - T')Y'(0)$, so that every generated sample has known potential outcomes and ground-truth individual treatment effect.

The key methodological idea is to treat causal properties as *design-time control primitives*. We specify three user-defined functions $\Psi = \{\alpha(X), \tau(X), \kappa(X, T)\}$, corresponding to overlap, treatment-effect heterogeneity, and unmeasured confounding.

Treatment-effect control. The target CATE function is $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, which controls both the magnitude and heterogeneity of treatment effects.

Unmeasured confounding control. We encode residual confounding beyond X as a treatment-arm-specific difference in potential outcomes. For binary treatment, define

$$\kappa(x, t) = t\Big(\mathbb{E}[Y(0) \mid X = x, T = 1] - \mathbb{E}[Y(0) \mid X = x, T = 0]\Big) + (1 - t)\Big(\mathbb{E}[Y(1) \mid X = x, T = 1] - \mathbb{E}[Y(1) \mid X = x, T = 0]\Big).$$

Thus, $\kappa(x, t)$ controls how strongly treatment assignment is associated with the counterfactual outcome not observed under treatment arm t, inducing departures from conditional ignorability; $\kappa \equiv 0$ corresponds to conditional ignorability in mean.

Overlap control. Overlap is specified through a covariate-dependent conditional density ratio, $\alpha(x) = \frac{p(X=x|T=0)}{p(X=x|T=1)}$, describing the similarity of the covariate distributions across treatment groups. By Bayes' rule, $\log \alpha(x) = -\operatorname{logit} e(x) + \log \frac{p(T=1)}{p(T=0)}$, so the density-ratio and propensity-score formulations differ by a treatment-prevalence term. We use the density-ratio form because it treats both arms symmetrically and can be computed from the mixed-type decoder likelihoods without an auxiliary propensity model; propensity-based diagnostics are reported as complementary checks (Section 3.7). Values near one indicate strong overlap, whereas large or small values indicate regions dominated by one treatment group.

Given these targets, we seek a generator whose induced causal quantities $\alpha_\theta(X)$, $\tau_\theta(X, T)$, and $\kappa_\theta(X, T)$ approximate $\alpha(X)$, $\tau(X)$, and $\kappa(X, T)$ while preserving the empirical distribution of (X, T, Y). We formulate this as a soft-constrained optimization problem:

$$\min_\theta \; d(P_\theta(X', T', Y'), P_{\text{obs}}(X, T, Y)) + \lambda_\alpha R_\alpha + \lambda_\tau R_\tau + \lambda_\kappa R_\kappa,$$

where $d(\cdot, \cdot)$ measures distributional discrepancy, R_α, R_τ, and R_κ penalize causal-target violations, and $\lambda_\alpha, \lambda_\tau, \lambda_\kappa$ control the realism–control trade-off (Section 3.4 gives the generator-specific form, with separate rigidity for the mean and variance components). If the specified causal targets are incompatible with the observed distribution, the objective returns the closest feasible generator under the chosen weights (Section 4.3).

3.2 CausalMix generator

CAUSALMIX factorizes the synthetic data-generating process as $p_\theta(X, T, Y) = p_\theta(T)\, p_\theta(X \mid T)\, p_\theta(Y(0), Y(1) \mid X, T)$. This factorization is a sampling parameterization, not a causal claim that treatment causes covariates. It is useful because conditioning the high-dimensional mixed-type

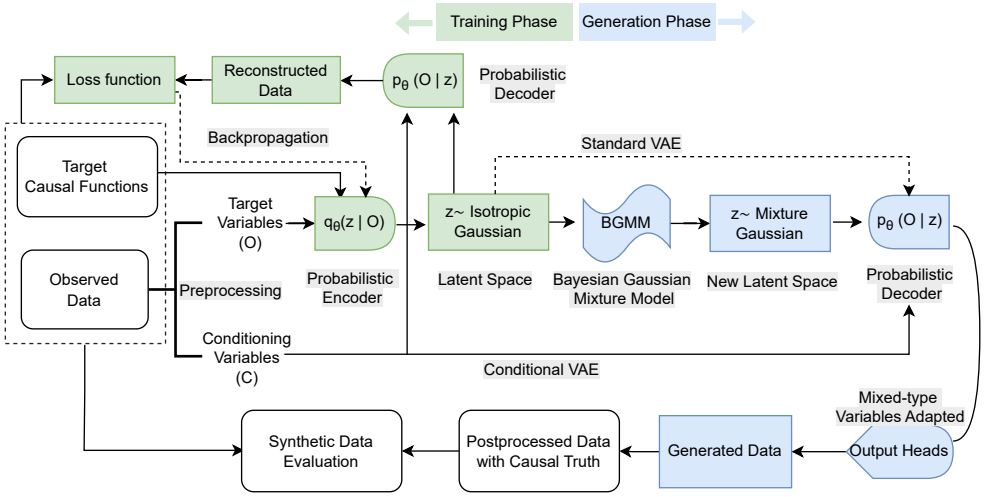


Figure 1: Overview of the CAUSALMIX generative framework.

covariate vector X on the low-dimensional treatment T stabilizes generation and enables direct overlap control through $p_\theta(X \mid T)$.

The generator consists of three modules: (i) an empirical Bernoulli treatment model $p_\theta(T)$; (ii) a pre-treatment generator $G_{X,\theta}$ for $X \mid T$; and (iii) a post-treatment generator $G_{Y,\theta}$ for $(Y(0), Y(1)) \mid X, T$. Both $G_{X,\theta}$ and $G_{Y,\theta}$ are implemented as conditional variational autoencoders (VAEs). For a target variable block O and conditioning variables C, the encoder maps O to a latent posterior $q_\theta(z \mid O)$, and the decoder generates O from the concatenated pair (z, C). In CAUSALMIX, $C = T$ for $G_{X,\theta}$ and $C = (X, T)$ for $G_{Y,\theta}$. Figure 1 provides an overview of the CAUSALMIX framework.

Mixed-type decoder heads. Real observational datasets contain continuous, binary, categorical, and integer-valued variables. To support such data without ad hoc discretization, each conditional VAE uses type-specific decoder heads: Bernoulli likelihoods for binary variables, categorical likelihoods for discrete variables, and Gaussian likelihoods for continuous variables. These likelihoods serve two roles: they define the reconstruction objective for realistic tabular generation, and they provide the feature-wise log-likelihoods needed to compute the mixed-type density ratio used in overlap control. Full likelihood formulas are given in Appendix B. Decoders output distribution parameters (Bernoulli or categorical logits, and Gaussian means with bounded log-variances); the causal losses below use only decoder means or probabilities.

$G_{X,\theta}$ and $G_{Y,\theta}$ are trained separately on the observed data. Synthetic data are then generated sequentially by sampling T', then $X' \mid T'$ from $G_{X,\theta}$, and finally $(Y'(0), Y'(1)) \mid X', T'$ from $G_{Y,\theta}$. Thus, the generated covariates serve as conditioning inputs to the outcome generator, yielding synthetic datasets with known individual and average treatment effects.

3.3 Causal-control losses

Let $m_{a,\theta}(X, T)$ denote the decoder mean of $Y(a)$ given (X, T), $a \in \{0, 1\}$. Because τ and κ are conditional-mean contrasts, the causal losses use decoder means rather than sampled outcomes. Expectations are minibatch averages.

Treatment-effect loss. The induced effect is $\tau_\theta(X, T) = m_{1,\theta}(X, T) - m_{0,\theta}(X, T)$, with residual $r_\tau = \tau_\theta(X, T) - \tau(X)$. We define

$$L_\tau^{\text{mean}} = \mathbb{E}\big[w_\tau\, r_\tau^2 + (1 - w_\tau)\,\text{SmoothL1}(r_\tau)\big]\,, \qquad L_\tau^{\text{var}} = \text{Var}(r_\tau),$$

where $w_\tau \in [0, 1]$ blends the quadratic and Smooth L1 terms. The quadratic term anchors mean alignment and the Smooth L1 term improves robustness to outliers. L_τ^{var} is the minibatch variance of the calibration residual, not a decoder variance: it penalizes uneven calibration across covariate regions and does not shrink the heterogeneity of τ itself.

Confounding loss. The induced confounding contrast compares decoder outputs under observed and flipped treatment assignments: $\kappa_\theta(X,T) = T\{m_{0,\theta}(X,T) - m_{0,\theta}(X,1-T)\} + (1-T)\{m_{1,\theta}(X,1-T) - m_{1,\theta}(X,T)\}$. Thus, confounding is realized through the dependence of the decoder means on T given X. With $r_\kappa = \kappa_\theta(X,T) - \kappa(X,T)$, we define L_κ^{mean} and L_κ^{var} analogously, with blend weight w_κ.

Overlap loss. For the pre-treatment generator, let $z \sim q_\theta(z \mid X)$ and $\Delta_\theta(X) = \log p_\theta(X \mid z, T = 0) - \log p_\theta(X \mid z, T = 1)$, where the same latent code is used for both treatment values and the log-likelihoods sum feature-wise decoder terms across continuous, binary, and categorical covariates (Appendix B). Given the target $\log \alpha(X)$, we define $L_\alpha = \mathbb{E}\big[\{\Delta_\theta(X) - \log \alpha(X)\}^2\big]$. This term shapes the decoder-level likelihood ratio and is the main mechanism for design-time control of covariate overlap; propensity-based diagnostics (Section 3.7) are complementary.

3.4 Unified training objective

Each conditional VAE is trained using the negative evidence lower bound, $L_{\text{VAE}} = -\mathbb{E}_{q_\theta(z|O)}\left[\log p_\theta(O \mid z, C)\right] + \lambda_{\text{KL}}\text{KL}(q_\theta(z \mid O) \,\|\, p(z))$, where O denotes the generated target variables and C the conditioning variables. Each generator has its own objective, combining fidelity with the causal penalties relevant to it:

$$L_X(\theta_X) = L_{\text{VAE}}^X + \lambda_\alpha L_\alpha, \tag{1}$$

$$L_Y(\theta_Y) = L_{\text{VAE}}^Y + \lambda_\tau L_\tau^{\text{mean}} + \lambda_\tau^{\text{var}} L_\tau^{\text{var}} + \lambda_\kappa L_\kappa^{\text{mean}} + \lambda_\kappa^{\text{var}} L_\kappa^{\text{var}}. \tag{2}$$

Setting all rigidity parameters to zero recovers a realism-only generator. Increasing them moves the solution along a realism–control frontier; it does not by itself reduce realism, and a trade-off appears when the requested causal targets are difficult to reconcile with the data or with one another (Section 4.3).

3.5 Mixture latent prior

Standard VAEs sample from an isotropic Gaussian prior, which can poorly match the multimodal latent geometry induced by real tabular data; mixture priors have been shown to improve sampling fidelity in tabular VAEs [21]. To improve sampling realism, CAUSALMIX fits a Bayesian Gaussian mixture model (BGMM) to encoder latent means after VAE training: $p_{\text{BGMM}}(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z; \mu_k, \Sigma_k)$. The Dirichlet-process prior encourages sparsity in mixture weights, allowing the number of active components to adapt to the data. At generation time, latent samples are drawn from $p_{\text{BGMM}}(z)$ and decoded conditionally. Fitting the BGMM post hoc, rather than as a trained prior, leaves the VAE and causal-control objectives unchanged and avoids extra mixture parameters and unstable component use during training, while improving realism for multimodal tabular distributions. We compare BGMM sampling against the standard Gaussian prior in experiments.

3.6 Optimization and practical tuning workflow

Models are trained with Adam using an 80/20 train–validation split and early stopping on validation loss. We log reconstruction, KL, treatment-effect, confounding, and overlap losses separately to diagnose whether distributional or causal terms dominate optimization. Optimization details are provided in Appendix C.

Tuning recipe. We recommend a staged workflow. (1) Fit a realism-only generator (all rigidity and variance weights set to zero). (2) Specify simple, domain-informed targets $\tau(X)$, $\kappa(X,T)$, and $\log \alpha(X)$. (3) Increase λ_τ,λ_κ and λ_α gradually, monitoring distributional fidelity and target-attainment errors. (4) Select an operating point where causal calibration improves without unacceptable loss of realism. (5) If needed, adjust the blend and variance weights: lower blend weights emphasize the Smooth L1 term, which limits the influence of large residuals, and larger variance weights can help avoid overly homogeneous generated effects when targets are low-dimensional or weakly nonlinear. In our experiments, $\lambda_\tau, \lambda_\kappa \approx 10^3$, $\lambda_\alpha \approx 10^1$–$10^2$, $w_\tau, w_\kappa \in [0.2, 0.4]$, and $\lambda_\tau^{\text{var}}, \lambda_\kappa^{\text{var}} \in [2\times10^{-2}, 2\times10^{-1}]$ gave stable operating regions; these are starting points, not universal defaults. Persistent target residuals despite larger rigidity and stable validation loss suggest that the requested regime may be incompatible with the data (Section 4.3).

3.7 Evaluation framework

Because CAUSALMIX is intended as a causal benchmarking sandbox, we evaluate generated data along three axes. First, *distributional fidelity* measures how closely synthetic samples match the observed data through marginal, pairwise, conditional, and joint metrics, including normalized Wasserstein distance, correlation and contingency similarity, conditional MMD, energy distance, and classifier two-sample tests. Second, *causal-structure fidelity* measures whether the generated data realize the user-specified causal controls, using CATE error and correlation, ATE error, confounding MAE and distributional distance, and overlap diagnostics based on decoder density ratios and propensity-score overlap. Third, *privacy diagnostics* assess record-level disclosure risk using distance-to-closest-record metrics, audited after training rather than being enforced. Full metric definitions are provided in Appendix D.

4 Experiments

We evaluate CAUSALMIX in four stages. First, we compare CAUSALMIX with CREDENCE [9], the closest prior generator that supports user-specified treatment effects and unmeasured confounding, on a public benchmark to assess performance in jointly achieving distributional realism and causal fidelity. Second, we validate whether CAUSALMIX can preserve realistic mixed-type clinical distributions while realizing user-specified treatment effects, unmeasured confounding, and overlap in a calibrated mCRPC cohort. Third, we examine the realism–control trade-off and causal control interactions through sensitivity analyses, cross-control effects, and stress tests with difficult or incompatible targets. Fourth, we use the validated generator as a synthetic ground-truth environment for benchmarking CATE estimators in a realistic comparative-safety setting, with brief demonstrations of additional applications, including hyperparameter tuning and power analysis for study design.

4.1 Public benchmark comparison with CREDENCE

Benchmark and design. We use the LaLonde temporary employment benchmark. The National Supported Work (NSW) program is a randomized trial evaluating the effect of a temporary employment program on post-intervention income [22]. Following standard observational benchmark practice, we combine NSW treated units with PSID control units [23]. The outcome is 1978 income, and pre-treatment covariates include age, education, race, degree status, and pre-intervention income.

Because CREDENCE does not support explicit overlap control or mixed-type likelihoods in the same way as CAUSALMIX, we restrict this comparison to shared axes: distributional realism, treatment-effect fidelity, and confounding fidelity. We specify a nonlinear treatment-effect target $\tau(X) = 0.15 + 0.10 \tanh(0.6\,\mathrm{education} + 0.5\,\mathrm{age} + 0.8\,\mathrm{nodegree})$, and use a constant unmeasured confounding target $\kappa(X, T) = 0.1$. We compare CREDENCE, CAUSALMIX with a Gaussian prior, and CAUSALMIX with the BGMM prior regarding causal and distributional fidelity of corresponding generated data, as described in Section 3.7.

Results. Table 1 summarizes representative metrics. Both CAUSALMIX variants outperform CREDENCE on causal-target fidelity. CATE/ITE MAE is substantially lower (0.007 for BGMM and 0.006 for Gaussian vs. 0.025 for CREDENCE), CATE correlation is markedly higher (0.978 for both vs. 0.298), and ATE error (0.004 and 0.002 vs. 0.006) and confounding MAE (0.036 and 0.029 vs. 0.069) are also lower.

CAUSALMIX-BGMM additionally provides substantially stronger distributional fidelity. Relative to CREDENCE, it achieves lower normalized Wasserstein distance (0.120 vs. 0.514) and normalized energy distance (0.011 vs. 0.254), together with higher correlation similarity (0.996 vs. 0.436) and C2ST complement (0.906 vs. 0.302). Qualitative diagnostics in Appendix F show the same pattern, with CAUSALMIX more closely preserving the reference marginal distributions and dependence structure (Figure 5 and Figure 6).

Within CAUSALMIX, BGMM substantially improves distributional fidelity over Gaussian sampling while maintaining comparably strong causal-target fidelity, supporting BGMM as the preferred sampling strategy when realistic data generation is a primary objective.

Table 1: Public benchmark comparison on NSW treated + PSID controls. Treatment-effect and confounding metrics assess fidelity to specified targets; distributional metrics assess fidelity to observed data. Values are mean (sd) over 10 generated datasets. Arrows indicate preferred direction. Full metrics are in Appendix F.

Axis	Metric	Direction	CM-BGMM	CM-Gauss	Credence
Treatment effect	CATE/ITE MAE	↓	0.007 (0.000)	0.006 (0.000)	0.025 (0.000)
Treatment effect	CATE correlation	↑	0.978 (0.000)	0.978 (0.002)	0.298 (0.017)
Treatment effect	ATE error	↓	0.004 (0.000)	0.002 (0.000)	0.006 (0.001)
Confounding	Confounding MAE	↓	0.036 (0.001)	0.029 (0.001)	0.069 (0.001)
Distribution	Normalized Wasserstein	↓	0.120 (0.001)	0.262 (0.005)	0.514 (0.008)
Distribution	Correlation similarity	↑	0.996 (0.001)	0.960 (0.009)	0.436 (0.000)
Distribution	Energy distance	↓	0.011 (0.000)	0.046 (0.002)	0.254 (0.009)
Distribution	C2ST complement	↑	0.906 (0.008)	0.595 (0.015)	0.302 (0.006)

CM = CAUSALMIX; Gauss = Gaussian prior; MAE = mean absolute error; C2ST = classifier two-sample test

4.2 Clinical-cohort generator validation

Clinical cohort. We next evaluate CAUSALMIX on a claims-based cohort from a comparative safety study of metastatic castration-resistant prostate cancer (mCRPC). The treatment is first-line abiraterone versus enzalutamide, and the outcome is treatment-related hospitalization or emergency department use. The cohort contains 4,098 patients and 18 baseline covariates of continuous, binary, and categorical types, including demographics, prior treatments, metastatic sites, comorbidities, and concomitant medications.

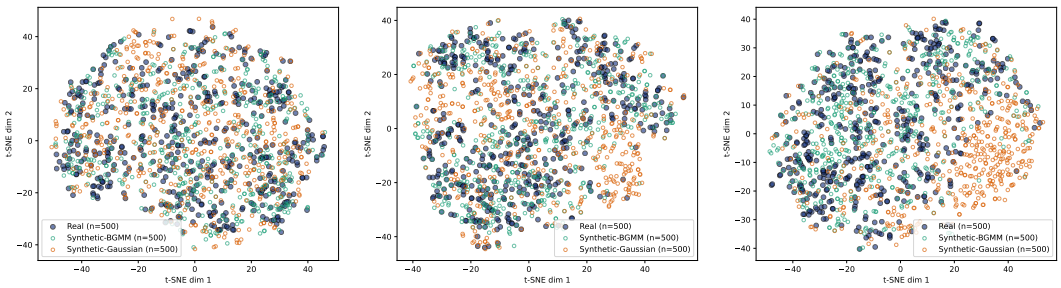
Causal scenarios. We define three progressively more challenging causal regimes. Scenario 1 specifies a constant treatment effect, no unmeasured confounding, and perfect overlap. Scenario 2 introduces linear treatment-effect heterogeneity, mild confounding, and moderate constant overlap. Scenario 3 combines nonlinear heterogeneous effects, covariate-dependent unmeasured confounding, and covariate-dependent overlap. These scenarios test whether the generator can jointly preserve empirical structure and realize user-specified causal controls.

Results. CAUSALMIX preserves realistic mixed-type distributions while realizing the specified causal controls in all scenarios. The benefit of the BGMM prior is most pronounced in Scenario 3, where multimodality and causal complexity are strongest. Compared with the Gaussian prior, BGMM substantially improves distributional fidelity: normalized energy distance decreases from 0.080 to 0.011, and marginal, pairwise, conditional, and joint metrics improve similarly (Appendix G). Figure 2 shows that BGMM sampling follows the geometry of the observed data more closely as causal complexity increases. Causal fidelity remains high: in Scenario 3, BGMM achieves CATE/ITE MAE of 0.004, CATE correlation of 0.956, confounding MAE of 0.004, and decoder-level overlap MSE of 0.004, whereas the Gaussian prior's overlap MSE is far larger (0.300). Figure 3 confirms close alignment between target and generated CATE structure.

Privacy diagnostics indicate a realism–privacy trade-off. The standard Gaussian prior provides stronger record-level separation but substantially lower fidelity, whereas BGMM achieves better fit with moderately reduced protection. Across scenarios, BGMM protection fractions remain above 0.5 and median distance ratios above one, with the largest relative reductions concentrated in the lower tail of the distance-ratio distribution. This pattern suggests localized proximity rather than systematic memorization. Cohort details, scenario definitions, and full distributional, causal, and privacy results are in Appendix G.

4.3 Realism–control trade-off and sensitivity to causal rigidity

One-at-a-time sweeps. We vary each rigidity parameter around its recommended operating value in mCRPC Scenario 3, holding the others fixed, so that the Realism–control trade-off is assessed in the full model rather than with other controls disabled. Increasing λ_α improves overlap calibration at a modest cost in realism around the default, while increasing λ_τ or λ_κ improves treatment-effect or confounding alignment with little change in global distributional fidelity (Appendix H).



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 2: 2D joint embedding of synthetic data under BGMM and Gaussian priors compared with the real mCRPC dataset across causal scenarios. Scenario 1: homogeneous effect, no confounding, perfect overlap; Scenario 2: linear heterogeneous effect, mild confounding, constant overlap; Scenario 3: nonlinear heterogeneity, covariate-dependent confounding and overlap.

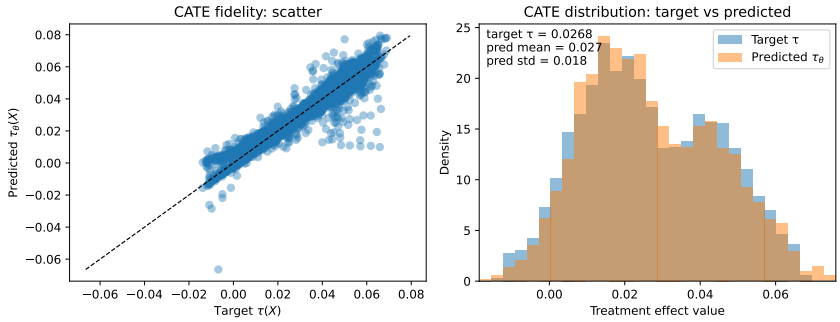


Figure 3: Treatment-effect fidelity in mCRPC Scenario 3 under CAUSALMIX-BGMM. The generated CATE aligns closely with the target function at both unit and distributional levels.

Cross-control effects. Because the treatment-effect and confounding losses share similar terms, we also check whether tuning one control affects the others. In nonlinear Scenario 3 with $\lambda_\alpha = 10$ and $\lambda_\kappa = 1000$, increasing λ_τ from approximately zero to 3000 reduces CATE MAE from 0.028 to 0.003 and raises CATE correlation from 0.42 to 0.97, while realism and the other causal errors stay small (Appendix H.2). When the target is compatible with the data, stronger control therefore need not cost realism or other causal properties.

Difficult or incompatible targets. Joint feasibility of the requested targets cannot be guaranteed a priori. When a target conflicts with the data distribution, the cost appears as residual target error, reduced fidelity, or both. In a simulation with known truth ($\tau(X) = 1 + 0.5X_1$, no unmeasured confounding), requesting increasingly incompatible effect functions leaves confounding and overlap errors small and CATE MAE low (0.018 to 0.042), but the most incompatible target raises energy distance from 0.017 to 0.050 and the treated-arm outcome Wasserstein distance from 0.13 to 1.39. In mCRPC Scenario 3, near-perfect overlap with a fourfold stronger confounding target is met (confounding MAE 0.205 to 0.003) at the cost of higher CATE MAE (0.003 to 0.016) (Appendix H.3). We therefore recommend assessing target attainment and fidelity jointly.

Optimization diagnostics. The constituent validation losses decrease and then stabilize, with no divergence among the competing objectives (Appendix F.4). Residual target errors at a given operating point therefore reflect the finite-weight trade-off rather than unconverged causal loss.

4.4 Benchmarking CATE estimators in a calibrated sandbox

After validating the generator, we use CAUSALMIX as a calibrated synthetic ground-truth environment for comparing CATE estimators in the heterogeneous treatment safety study of mCRPC. This experiment demonstrates the intended use case: evaluating estimators under a distribution resembling

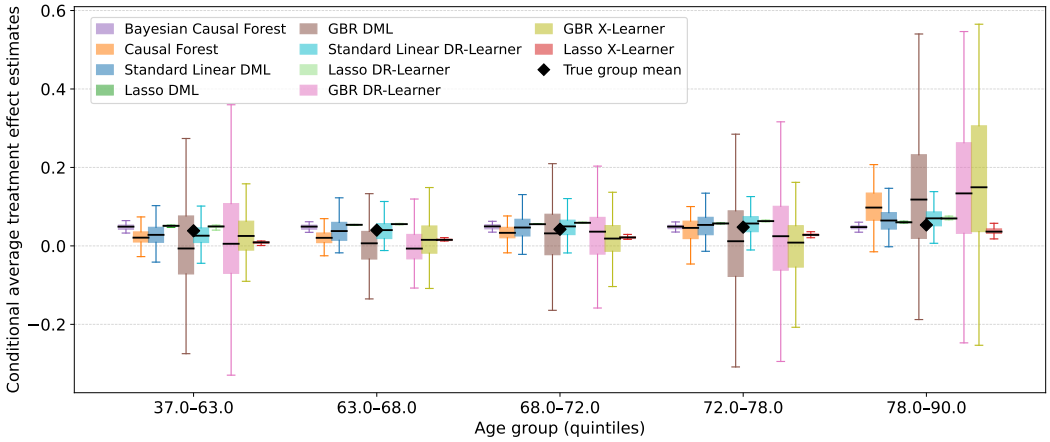


Figure 4: Grouped CATE estimates by age quintile. For each age group, boxplots summarize individual-level CATE estimates from each estimator in the subgroup with Charlson = 0, no prior CVD, under a calibrated synthetic data-generating process without unmeasured confounding ($\kappa = 0$). Black diamonds indicate the true group-level mean CATE.

the target application while retaining known individual treatment effects and controlled violations of causal assumptions.

Design. We generate synthetic mCRPC datasets under a regime with linear treatment-effect heterogeneity, moderate overlap, and either no unmeasured confounding or moderate unmeasured confounding. The target treatment effect is $\tau(X) = 0.05 + 0.015\,\mathrm{cvd} + 0.01\,\mathrm{age} + 0.01\,\mathrm{Charlson}$. The two confounding settings are $\kappa(X, T) = 0$ and $\kappa(X, T) = 0.02$. For each setting, we generate 50 independent synthetic datasets and retain the true ATE and individual treatment effects.

Estimators and metrics. We compare representative estimators from four families: meta-learners [24], doubly robust (DR) learners [25, 26], double machine learning (DML) estimators [27], and tree-based methods. Specifically, we include X-learners, DR learners, DML estimators, causal forests [28], and Bayesian causal forests (BCF) [29]. Since the synthetic data contain ground-truth potential outcomes, we evaluate population-level and heterogeneous-effect performance using ATE bias/RMSE, ATE coverage, PEHE (CATE RMSE), CATE coverage, and runtime. Implementation details are provided in Appendix I.

Results. Without unmeasured confounding, most estimators achieve low ATE error, but they differ substantially in CATE uncertainty calibration. Bayesian causal forests and the standard linear DR learner achieve the highest CATE coverage, while lasso-based DML and DR learners under-cover despite competitive ATE accuracy. Figure 4 summarizes age-stratified CATE estimates in a clinically relevant subgroup; Bayesian causal forests track the true group-level CATE pattern most closely, whereas several regularized learners oversmooth heterogeneous effects.

Introducing unmeasured confounding shifts ATE estimates by approximately the injected magnitude (0.02) for all estimators. For X-learners, whose ATE bias is negative without confounding, this shift offsets the existing bias and lowers ATE RMSE, which should not be read as robustness. Relative CATE coverage patterns persist, showing that estimator choice still matters for heterogeneity estimation and uncertainty quantification even when no estimator can eliminate hidden confounding. This illustrates the practical value of a calibrated causal sandbox: it separates ATE accuracy, heterogeneous-effect recovery, uncertainty calibration, and computational cost under controlled but realistic data-generating regimes.

4.5 Additional downstream applications

Using the calibrated no-confounding mCRPC generator as a ground-truth environment, we demonstrate two further applications (details in Appendix J).

Estimator tuning. We tune causal-forest hyperparameters over minimum leaf size (5–80) and number of trees (500–4000), with 50 synthetic datasets per configuration. ATE accuracy is insensitive to the grid (RMSE about 0.009), whereas larger leaves lower PEHE (0.036 to 0.023) at the cost of CATE coverage (about 0.9 to 0.75); the number of trees mainly affects runtime. Once the number of trees is moderately large, leaf size is more consequential than tree count, and leaf sizes of 20–40 with 1,000–2,000 trees balance accuracy, calibration, and runtime.

Power analysis for study design. To plan an observational claims-based study of effect modification by baseline cardiovascular disease, we estimate the power of a best-linear-projection test on causal forests (minimum leaf size 80, chosen to stabilize modifier signals) over 100 synthetic datasets per sample size. Power is 0.13 at $n = 200$, 0.54 at $n = 500$, 0.67 at $n = 1{,}000$, and 0.94 at $n = 2{,}000$, approaching 1.0 at $n = 5{,}000$, whereas the forest ranks the true modifier first by variable importance only at $n = 5{,}000$. Detecting effect modification thus requires fewer patients than reliably identifying the modifier.

5 Discussion and Conclusion

Summary. We introduced CAUSALMIX, a controllable generative sandbox for causal inference with realistic mixed-type tabular data, with a unified evaluation suite spanning distributional fidelity, causal fidelity, and privacy diagnostics. It treats treatment effects, unmeasured confounding, and overlap as explicit design-time controls while anchoring generation to an observed empirical distribution. The generated synthetic datasets combine realistic data structure with known counterfactual ground truth, addressing complementary limitations of real observational data and hand-designed simulators.

Usefulness and empirical findings. CAUSALMIX is intended for researchers who have an observational dataset of interest and need a calibrated environment for benchmarking estimators, stress-testing assumptions, or planning studies under plausible causal regimes. It is not meant to replace mechanistic simulators when reliable domain-specific models exist, but offers a practical alternative when they are unavailable or hard to calibrate. On the public benchmark, CAUSALMIX, especially with the BGMM prior, preserved the observed distribution far better than CREDENCE while matching the specified treatment-effect and confounding targets more closely. In the mCRPC case study, the validated generator revealed differences among CATE estimators that would be hard to assess from real data alone: estimators with similar ATE accuracy differed in CATE calibration, and injected unmeasured confounding shifted ATE estimates by about the injected magnitude for all methods. The same generator supported hyperparameter tuning and power analysis for study design.

Limitations. As with any learned generator, strong aggregate fidelity metrics do not guarantee preservation of rare events, extreme covariate patterns, or higher-order dependencies. Causal controls are user-specified, so poorly chosen functions may encode incorrect assumptions. Joint feasibility of the requested targets also cannot be guaranteed a priori and is revealed through the realism–control frontier (Section 4.3). Rigidity parameters require dataset- and target-dependent tuning. There is no automatic selection rule, but we provide a tuning workflow and starting ranges (Section 3.6). We therefore view CAUSALMIX as a sensitivity-analysis tool rather than a method for discovering the true causal DGP, and recommend verifying the realized causal structure with the proposed diagnostics.

Future work and conclusion. Each causal regime currently requires retraining. This is a modest cost for the intended offline use, where a limited number of scientifically motivated regimes are examined and many datasets are then drawn from each fitted model. Amortizing control across regimes, for example by conditioning the generator on an encoding of the target functions, would mainly benefit interactive or large-scale sweeps, and would need evaluation against separately trained models. Extensions to longitudinal treatments, time-varying confounding, missing data, and richer outcome processes would further broaden applicability. Overall, CAUSALMIX makes the realism–control trade-off explicit and measurable, supporting more informative evaluation of causal estimators when real data alone cannot provide causal ground truth.

Acknowledgments and Disclosure of Funding

We thank Fang He (Amgen) for motivating discussions and help setting up the computing clusters. Q.Z. was supported by an Amgen predoctoral fellowship, and part of this work was performed using Amgen computational resources. R.N. was supported by the National Science Foundation under Grant No. 2502299.

Competing interests (outside the submitted work). Q.Z. reports personal fees from Epidemiologic Research & Methods, LLC. T.L.L. reports consulting fees and travel support from Amgen Inc. and Merck & Co., Inc. for methodologic consulting services.

AI use disclosure. Generative AI tools (including DeepSeek, ChatGPT and Claude) were used to assist with language editing, manuscript refinement and code development. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

[1] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[2] P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.

[3] Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. Empirical analysis of model selection for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.01939*, 2022.

[4] Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*, 23:2892–2910, 2024.

[5] Sarah Friedrich and Tim Friede. On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, 66(1):2200212, 2024.

[6] Juan L Gamella. *Physical Testbeds for AI Methodology*. PhD thesis, ETH Zurich, 2025.

[7] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

[8] Susan Athey, Guido W Imbens, Jonas Metzger, and Evan Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 240(2):105076, 2024.

[9] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *International conference on machine learning*, pages 17346–17358. PMLR, 2022.

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

[11] Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

[12] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[13] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15883–15894. Curran Associates, Inc., 2021.

[14] Jessica M Franklin, Sebastian Schneeweiss, Jennifer M Polinski, and Jeremy A Rassen. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis*, 72:219–226, 2014.

[15] Shao-Hsien Liu, Stavroula A Chrysanthopoulou, Qiuzhi Chang, Jacob N Hunnicutt, and Kate L Lapane. Missing data in marginal structural models: a plasmode simulation study comparing multiple imputation and inverse probability weighting. *Medical care*, 57(3):237–243, 2019.

[16] Vanessa Ress and Eva-Maria Wild. Comparing methods for estimating causal treatment effects of administrative health data: A plasmode simulation study. *Health Economics*, 33(12):2757–2777, 2024.

[17] Raymond B Cattell and Joseph Jaspers. A general plasmode (no. 30-10-5-2) for factor analytic exercises and research. *Multivariate behavioral research monographs*, 1967.

[18] Gary L Gadbury, Qinfang Xiang, Lin Yang, Stephen Barnes, Grier P Page, and David B Allison. Evaluating statistical methods using plasmode data sets in the age of massive public databases: an illustration using false discovery rates. *PLoS genetics*, 4(6):e1000098, 2008.

[19] Laura K Vaughan, Jasmin Divers, Miguel A Padilla, David T Redden, Hemant K Tiwari, Daniel Pomp, and David B Allison. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Computational statistics & data analysis*, 53(5):1755–1766, 2009.

[20] Pamela A Shaw, Susan Gruber, Brian D Williamson, Rishi Desai, Susan M Shortreed, Chloe Krakauer, Jennifer C Nelson, and Mark J van der Laan. A cautionary note for plasmode simulation studies in the setting of causal inference. *arXiv preprint arXiv:2504.11740*, 2025.

[21] Patricia A Apellániz, Juan Parras, and Santiago Zazo. An improved tabular data generator with vae-gmm integration. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1886–1890. IEEE, 2024.

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

[23] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

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

[25] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[26] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[27] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[28] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

[29] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[30] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[31] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[32] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[33] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

[34] Henry F Inman and Edwin L Bradley Jr. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, 18(10):3851–3874, 1989.

[35] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[36] Amy Steier, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter. Synthetic data privacy metrics. *arXiv preprint arXiv:2501.03941*, 2025.

[37] Ivan Henriquez, Mack Roach III, Todd M Morgan, Alberto Bossi, Junior A Gomez, Oscar Abuchaibe, and Felipe Counago. Current and emerging therapies for metastatic castration-resistant prostate cancer (mcrpc). *Biomedicines*, 9(9):1247, 2021.

[38] Martin W Schoen, Kenneth R Carson, Seth A Eisen, Charles L Bennett, Suhong Luo, Melissa A Reimers, Eric M Knoche, Alison L Whitmer, Yan Yan, Bettina F Drake, et al. Survival of veterans treated with enzalutamide and abiraterone for metastatic castrate resistant prostate cancer based on comorbid diseases. *Prostate Cancer and Prostatic Diseases*, 26(4):743–750, 2023.

[39] Stephen J Freedland, Xuehua Ke, Marie-Hélène Lafeuille, Hela Romdhani, Frederic Kinkead, Patrick Lefebvre, Allison Petrilla, Zul Pulungan, Seung Kim, Denise M D'Andrea, et al. Identification of patients with metastatic castration-sensitive or metastatic castration-resistant prostate cancer using administrative health claims and laboratory data. *Current Medical Research and Opinion*, 37(4):609–622, 2021.

[40] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, pages 443–470, 2013.

A Comparison with Existing Approaches

Table 2 summarizes how existing approaches to synthetic data for causal inference differ in empirical realism, control over treatment effects, confounding, and overlap, and support for mixed-type variables. Entries describe each approach class, not every method within it.

Table 2: Qualitative comparison of approaches to synthetic data for causal inference.

Approach	Data-anchored	Effect control	Confounding control	Overlap control	Mixed-type
Parametric simulators	Limited	Yes	Yes	Often	Limited
Semi-synthetic benchmarks	Covariates only	Fixed	Fixed or none	Indirect	Inherited
Data-fit neural generators	High	Varies	Varies	Rarely explicit	Often
CREDENCE	Yes	Yes	Yes	No	Limited
CAUSALMIX	Yes	Yes	Yes	Yes	Yes

B Mixed-Type Likelihoods and Log-Density Ratios

This appendix gives the likelihood formulas used for mixed-type variables and the induced log-density ratio used for overlap control.

For the pre-treatment generator, the decoder defines

$$p_\theta(X \mid z, T = t) = \prod_{j=1}^{V} p_\theta(X_j \mid z, T = t),$$

so that

$$\log p_\theta(X \mid z, T = t) = \sum_{j=1}^{V} \log p_\theta(X_j \mid z, T = t).$$

The same feature-wise likelihoods are used for reconstruction and for computing the overlap-control term.

B.1 Binary variables

For a binary variable $X_j \in \{0, 1\}$, the decoder outputs a logit s_j and probability

$$p_j = \sigma(s_j) = \frac{1}{1 + \exp(-s_j)}.$$

The Bernoulli log-likelihood is

$$\log p_\theta(X_j \mid z, T = t) = X_j \log p_j + (1 - X_j) \log(1 - p_j).$$

B.2 Categorical variables

For a categorical variable $X_j \in \{0, \ldots, K - 1\}$, the decoder outputs logits $s_j \in \mathbb{R}^K$, with

$$p_{j,c} = \frac{\exp(s_{j,c})}{\sum_{\ell=1}^{K} \exp(s_{j,\ell})}.$$

The log-likelihood for observed class X_j is

$$\log p_\theta(X_j \mid z, T = t) = \log p_{j,X_j} = s_{j,X_j} - \log\left(\sum_{\ell=1}^{K} \exp(s_{j,\ell})\right).$$

B.3 Continuous variables

For a continuous variable $X_j \in \mathbb{R}$, the decoder outputs a mean μ_j and bounded log-variance $\log \sigma_j^2$, assuming

$$X_j \mid z, T = t \sim \mathcal{N}(\mu_j, \sigma_j^2).$$

The Gaussian log-density is

$$\log p_\theta(X_j \mid z, T = t) = -\frac{1}{2}\left[\log(2\pi) + \log \sigma_j^2 + \frac{(X_j - \mu_j)^2}{\sigma_j^2}\right].$$

B.4 Log-density ratio for overlap control

The conditional log-likelihood for treatment arm t is

$$\log p_\theta(X \mid T = t) = \sum_{j=1}^{V} \log p_\theta(X_j \mid z, T = t).$$

The induced log-density ratio is

$$\Delta_\theta(X) = \log p_\theta(X \mid T = 0) - \log p_\theta(X \mid T = 1).$$

Given a target $\log \alpha(X)$, the overlap penalty is

$$L_\alpha = \mathbb{E}_X\left[\{\Delta_\theta(X) - \log \alpha(X)\}^2\right].$$

This mechanism allows the generator to shape covariate overlap using likelihoods appropriate for continuous, binary, and categorical covariates.

C Optimization Details

All conditional VAE modules are trained using Adam with learning rate 10^{-3} and mini-batches drawn from an 80/20 train–validation split. Training uses early stopping on validation loss. During training, the reconstruction term, KL term, and individual causal penalties are logged separately.

C.1 Forward pass

For each mini-batch, the encoder outputs $(\mu, \log \sigma^2)$ and latent samples are drawn using the reparameterization trick,

$$z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

The latent sample is concatenated with conditioning variables before decoding. For the pre-treatment generator, the conditioning variable is T. For the post-treatment generator, the conditioning variables are (X, T).

C.2 Post-treatment causal penalties

The post-treatment generator decodes both potential outcomes $Y(0)$ and $Y(1)$. Reconstruction is evaluated only on the factual outcome corresponding to the observed treatment arm, while treatment-effect and confounding penalties use both potential-outcome heads.

To compute the confounding penalty, the decoder is evaluated twice: once under observed treatment T and once under flipped treatment $1 - T$, using the same covariates. This yields the induced confounding contrast $\kappa_\theta(X, T)$ used in the main objective.

C.3 BGMM fitting

After VAE training, the encoder latent means $\{\mu_\theta(O_i)\}_{i=1}^n$ are collected and used to fit a Bayesian Gaussian mixture model. The fitted mixture defines the sampling prior

$$p_{\mathrm{BGMM}}(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z; \mu_k, \Sigma_k).$$

The BGMM is fit post hoc and does not participate in gradient-based optimization. We compare BGMM sampling with standard Gaussian sampling in the experiments.

D Evaluation Metrics

We evaluate synthetic data along three axes: distributional fidelity, causal-structure fidelity, and privacy. All metrics are computed by comparing synthetic data to the observed data and, where applicable, to the user-specified causal targets. A summary of evaluation metrics is provided in Table 3.

Table 3: **Summary of evaluation metrics across three axes: distributional fidelity, causal-structure fidelity, and privacy.**

Axis	Category (Var Types)	Metric	Description / Formula
Distributional Fidelity	Marginal (cont.)	Normalized Wasserstein	$\frac{W_1(F^{(c)}_{\text{real}}, F^{(c)}_{\text{synth}})}{\text{sd}^{(c)}_{\text{real}} + \varepsilon}$.
	Marginal (cont.)	KSComplement (`sdmetrics`)	$1 - D_{\text{KS}}$.
	Marginal (disc.)	TVComplement (`sdmetrics`)	$1 - \frac{1}{2}\sum_x \lvert p_{\text{real}}(x) - p_{\text{synth}}(x)\rvert$.
	Pairwise (all vars)	SU Similarity	$1 - \lvert \text{SU}_{\text{real}}(i,j) - \text{SU}_{\text{synth}}(i,j)\rvert$.
	Pairwise (cont.–cont.)	CorrelationSimilarity (`sdmetrics`)	Similarity of Pearson correlations over cont. pairs.
	Pairwise (disc.–disc.)	ContingencySimilarity (`sdmetrics`)	Similarity of disc. contingency structure via total variation.
	Conditional ($Z \mid C$)	Weighted / Normalized MMD^2	Stratum-size weighted mixed-kernel $\text{MMD}^2(Z \mid C)$; normalized by within-real baseline.
	Joint (all vars)	Energy Dist.	Normalized energy distance.
	Joint (all vars)	C2ST	$1 - 2\lvert\text{AUC} - 0.5\rvert$.
Causal-Structure Fidelity	Treatment Effect	ATE Error	$\lvert\text{ATE}_\theta - \text{ATE}_{\text{target}}\rvert$.
	Treatment Effect	CATE MAE	$\mathbb{E}\big[\,\lvert\tau_\theta(X') - \tau(X')\rvert\,\big]$.
	Treatment Effect	CATE Correlation	Pearson correlation $\rho(\tau_\theta, \tau)$.
	Treatment Effect	Wasserstein Dist.	W_1 between empirical distributions of $\tau_\theta(X')$ and $\tau(X')$.
	Confounding	MAE (overall / by group)	$\mathbb{E}[\lvert\kappa_\theta(X', T') - \kappa(X', T')\rvert]$; by treatment arm.
	Confounding	Wasserstein Dist.	W_1 between empirical confounding distributions.
	Overlap	MSE	$\mathbb{E}\big[(\log \alpha_\theta(X') - \log \alpha(X'))^2\big]$.
	Overlap	Overlap Coefficient	$\int_0^1 \min\big(p_0(e), p_1(e)\big)\, de$.
	Overlap	Propensity AUC	AUC of logistic regression predicting T' from X'.
Privacy	DCR	Protection Fraction	$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{d_{\text{syn}}(i) > d_{\text{real}}(i)\}$.
	DCR	Distance Ratio	$\rho_i = \frac{d_{\text{syn}}(i)}{d_{\text{real}}(i) + \varepsilon}$.
	DCR	DCRBaselineProtection (`sdmetrics`)	$\min\Big(1,\ \widetilde{d}_{\text{synth}}/\widetilde{d}_{\text{rand}}\Big)$.

Abbreviations and notation: cont. = continuous; disc. = discrete (binary or categorical); TE = treatment effect; dist. = distance; C = conditioning variable; Z = all variables except C. For continuous feature c: $F^{(c)}_{\text{real}}, F^{(c)}_{\text{synth}}$ = empirical CDFs; W_1 = 1-Wasserstein distance; D_{KS} = Kolmogorov–Smirnov statistic. For discrete x: $p_{\text{real}}(x), p_{\text{synth}}(x)$ = empirical PMFs. $\text{SU}_{\text{real}}(i,j), \text{SU}_{\text{synth}}(i,j)$ = symmetric uncertainty (real vs. synthetic). MMD^2 = maximum mean discrepancy. (X', T') = synthetic samples; $\text{ATE}_{\text{target}} = \mathbb{E}[\tau(X')]$, $\text{ATE}_\theta = \mathbb{E}[\tau_\theta(X')]$. $e = \mathbb{P}(T' = 1 \mid X')$; $p_0(e), p_1(e)$ = group-specific densities. DCR = distance to closest record (encoded space); $d_{\text{syn}}(i), d_{\text{real}}(i)$ = NN distances to synthetic and to another real record; $\widetilde{d}_{\text{synth}}, \widetilde{d}_{\text{rand}}$ = median DCRs (synthetic, random). Metrics marked `sdmetrics` use the `sdmetrics` library.

D.1 Distributional fidelity

Marginal fidelity. For each continuous feature c, we compute normalized Wasserstein distance

$$W_{\text{norm}}(c) = \frac{W_1(F^{(c)}_{\text{real}}, F^{(c)}_{\text{synth}})}{\text{sd}^{(c)}_{\text{real}} + \varepsilon},$$

where W_1 is the 1-Wasserstein distance and $\varepsilon > 0$ is a numerical stabilizer. We also report KSComplement,

$$1 - D_{\mathrm{KS}},$$

where D_{KS} is the Kolmogorov–Smirnov statistic. For binary and categorical variables, we compute TVComplement,

$$\mathrm{TVComplement}(c) = 1 - \frac{1}{2} \sum_{x} \left| p_{\mathrm{real}}(x) - p_{\mathrm{synth}}(x) \right|.$$

Pairwise dependence. For continuous–continuous pairs, we report correlation similarity using Pearson correlations. For discrete–discrete pairs, we report contingency similarity based on total variation distance. We also compute symmetric uncertainty (SU) [30],

$$\mathrm{SU}(i,j) = \frac{2\,\mathrm{MI}(i,j)}{H(i) + H(j)}.$$

We report the similarity score

$$1 - \left| \mathrm{SU}_{\mathrm{real}}(i,j) - \mathrm{SU}_{\mathrm{synth}}(i,j) \right|.$$

Conditional fidelity. For a conditioning variable C, let Z denote all variables except C. We compare $p(Z \mid C)$ between real and synthetic data by stratifying on C and computing a stratum-weighted mixed-kernel MMD^2[31], using RBF kernels for continuous variables and Hamming kernels for discrete variables. We also report a normalized version using a within-real split as reference.

Joint fidelity. Global distributional fidelity is measured using normalized energy distance [32] and a classifier two-sample test (C2ST) [33]. The C2ST trains a classifier to distinguish real from synthetic samples and reports

$$1 - 2|\mathrm{AUC} - 0.5|,$$

which is maximized when real and synthetic samples are indistinguishable.

D.2 Causal-structure fidelity

Treatment-effect diagnostics. For each generated sample, the post-treatment generator produces potential outcomes $Y'(0)$ and $Y'(1)$, inducing

$$\tau_\theta(X') = Y'(1) - Y'(0).$$

Given target $\tau(X')$, we compute

$$\mathrm{MAE}_\tau = \mathbb{E}\left[\left| \tau_\theta(X') - \tau(X') \right| \right],$$

the Pearson correlation $\rho(\tau_\theta, \tau)$, ATE error

$$\left| \mathbb{E}[\tau_\theta(X')] - \mathbb{E}[\tau(X')] \right|,$$

and the Wasserstein distance between the distributions of $\tau_\theta(X')$ and $\tau(X')$.

Unmeasured confounding diagnostics. The generated confounding contrast is $\kappa_\theta(X', T')$, as defined in the main text. Given target $\kappa(X', T')$, we compute

$$\mathrm{MAE}_\kappa = \mathbb{E}\left[\left| \kappa_\theta(X', T') - \kappa(X', T') \right| \right],$$

group-wise MAE within $T' = 0$ and $T' = 1$, and Wasserstein distance between the distributions of $\kappa_\theta(X', T')$ and $\kappa(X', T')$.

Overlap diagnostics. For decoder-based overlap, we compare the generated log-density ratio

$$\Delta_\theta(X') = \log p_\theta(X' \mid T' = 0) - \log p_\theta(X' \mid T' = 1)$$

against the target $\log \alpha(X')$, reporting MSE and the fraction of samples within a user-specified tolerance band.

We also compute a model-agnostic propensity-based diagnostic. A logistic regression estimates

$$e(X') = \mathbb{P}(T' = 1 \mid X'),$$

and smooth densities of $e(X')$ are estimated separately for treated and control units. The overlap coefficient [34] is

$$\mathrm{OverlapCoeff} = \int_0^1 \min\{p_{T'=0}(e), p_{T'=1}(e)\}\, de.$$

Values closer to one indicate stronger common support.

E Privacy Evaluation

Privacy is evaluated using distance-to-closest-record (DCR) diagnostics [35, 36], which quantify whether synthetic records are unusually close to real records. Let $\mathcal{R} = \{r_i\}_{i=1}^{n}$ and $\mathcal{S} = \{s_j\}_{j=1}^{m}$ denote encoded real and synthetic datasets. For each real record r_i, define

$$d_{\text{syn}}(i) = \min_{s_j \in \mathcal{S}} \|r_i - s_j\|, \qquad d_{\text{real}}(i) = \min_{\substack{r_k \in \mathcal{R} \\ k \neq i}} \|r_i - r_k\|.$$

A real record is considered protected if its nearest synthetic neighbor is farther than its nearest other real neighbor:

$$\mathbf{1}_{\text{protected}}(i) = \mathbf{1}\{d_{\text{syn}}(i) > d_{\text{real}}(i)\}.$$

The DCR protection fraction is

$$\text{DCRProtection} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\text{protected}}(i).$$

We also compute the distance ratio

$$\rho(i) = \frac{d_{\text{syn}}(i)}{d_{\text{real}}(i) + \varepsilon},$$

and report summary statistics such as mean, median, and selected quantiles. Larger values indicate stronger record-level separation between real and synthetic data.

When available, we additionally report `DCRBaselineProtection`, which normalizes the median synthetic–real DCR by a random-data baseline and clips the score to $[0, 1]$.

F Public Benchmark Comparison Details

F.1 Benchmark construction

We use the LaLonde temporary employment benchmark. The National Supported Work (NSW) program is a randomized trial evaluating the effect of a temporary employment program on post-intervention income [22]. Following standard observational benchmark practice, we combine treated units from NSW with control units from the PSID comparison sample [23]. The outcome is 1978 income. Pre-treatment covariates include age, education, race indicators, degree status, marital status, and pre-intervention income.

Continuous variables are standardized before model fitting. Income variables are transformed consistently across methods when needed for numerical stability. We compare CAUSALMIX-BGMM, CAUSALMIX-Gaussian, and CREDENCE. Since CREDENCE does not support explicit overlap control in the same way, overlap-specific metrics are omitted from this comparison.

F.2 Target functions

The treatment-effect target is

$$\tau(X) = 0.15 + 0.10 \tanh(0.6\,\mathrm{education} + 0.5\,\mathrm{age} + 0.8\,\mathrm{nodegree}),$$

and the confounding target is constant:

$$\kappa(X, T) = 0.1.$$

We generate 10 synthetic datasets from each fitted generator and report mean and standard deviation across generated datasets.

F.3 Full quantitative results

Tables 4 and 5 report full causal-structure and distributional fidelity metrics for the public benchmark. Figures 5 and 6 provide qualitative diagnostics of dependence structure and marginal distributions.

Table 4: Causal fidelity on the LaLonde benchmark. Values are mean (sd) over 10 generated datasets.

Category	Metric	Direction	BGMM	Gaussian	Credence
Treatment Effect	CATE/ITE MAE	↓ better	0.007 (0.000)	0.006 (0.000)	0.025 (0.000)
Treatment Effect	CATE Correlation	↑ better	0.978 (0.000)	0.978 (0.002)	0.298 (0.017)
Treatment Effect	ATE Error	↓ better	0.004 (0.000)	0.002 (0.000)	0.006 (0.001)
Treatment Effect	TE Distribution Distance (W1)	↓ better	0.004 (0.000)	0.003 (0.000)	0.021 (0.000)
Confounding	Confounding MAE	↓ better	0.036 (0.001)	0.029 (0.001)	0.069 (0.001)
Confounding	Group-wise MAE (T=0)	↓ better	0.036 (0.000)	0.028 (0.000)	0.070 (0.001)
Confounding	Group-wise MAE (T=1)	↓ better	0.041 (0.001)	0.032 (0.001)	0.068 (0.001)
Confounding	Confounding Dist. (W1)	↓ better	0.036 (0.001)	0.029 (0.001)	0.069 (0.001)

Table 5: Distributional fidelity on the LaLonde benchmark. Values are mean (sd) over 10 generated datasets.

Category	Metric	Direction	BGMM	Gaussian	Credence
Marginal (cont.)	Normalized Wasserstein (mean)	↓ better	0.120 (0.001)	0.262 (0.005)	0.514 (0.008)
Marginal (cont.)	KSComplement (mean)	↑ better	0.894 (0.001)	0.850 (0.005)	0.643 (0.003)
Marginal (disc.)	TVComplement (mean)	↑ better	0.984 (0.001)	0.922 (0.002)	0.790 (0.010)
Pairwise (cont–cont.)	CorrelationSimilarity	↑ better	0.996 (0.001)	0.960 (0.009)	0.436 (0.000)
Pairwise (all vars)	SU similarity (mean)	↑ better	0.651 (0.013)	0.661 (0.015)	0.568 (0.001)
Pairwise (disc–disc.)	ContingencySimilarity (mean)	↑ better	0.943 (0.002)	0.843 (0.004)	0.662 (0.012)
Joint (all vars)	Normalized Energy Distance	↓ better	0.011 (0.000)	0.046 (0.002)	0.254 (0.009)
Joint (all vars)	C2ST (AUC complement)	↑ better	0.906 (0.008)	0.595 (0.015)	0.302 (0.006)

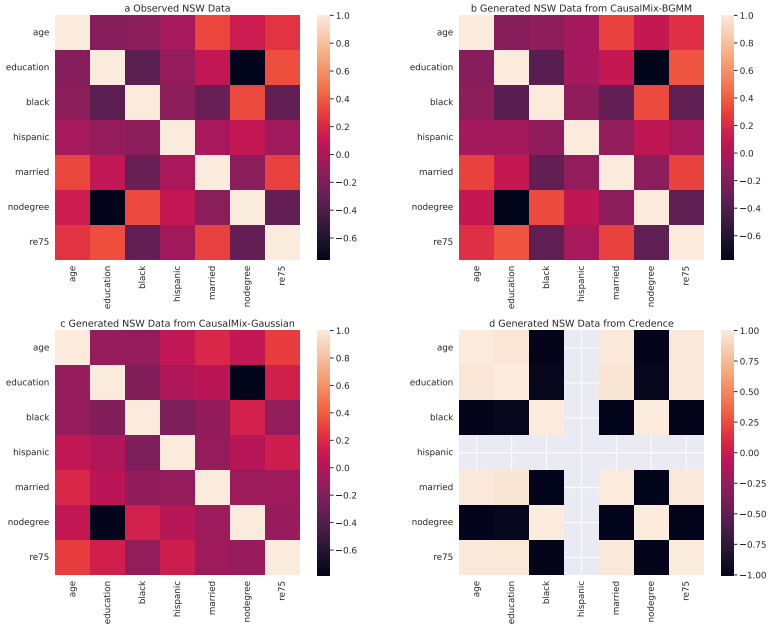


Figure 5: Correlation-structure comparison on the LaLonde benchmark. Heatmaps show pairwise correlations for the observed NSW-treated + PSID-control data and synthetic datasets generated by CAUSALMIX-BGMM, CAUSALMIX-Gaussian, and CREDENCE. CAUSALMIX, especially with the BGMM prior, better preserves the observed dependence structure, whereas CREDENCE exhibits near-degenerate block correlations.

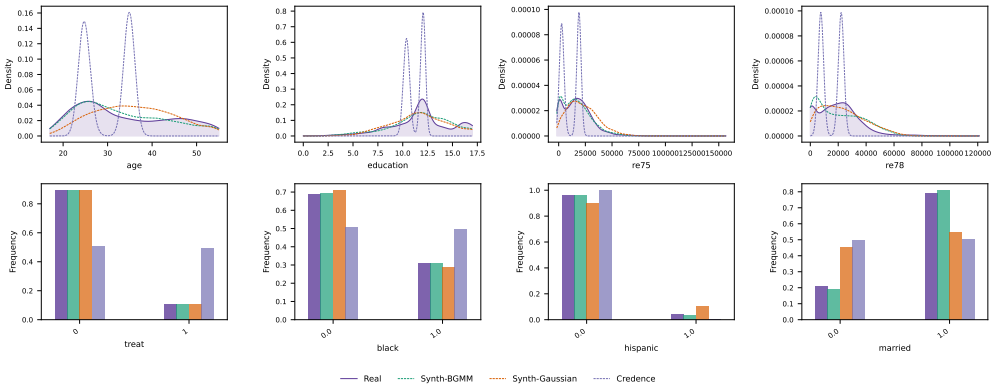


Figure 6: Marginal and categorical distribution comparison on the LaLonde benchmark. For continuous variables, curves show estimated densities; for binary variables, bars show empirical frequencies. Colors and line styles distinguish observed data, CAUSALMIX-BGMM, CAUSALMIX-Gaussian, and CREDENCE, as indicated by the legend. CAUSALMIX more closely tracks the observed marginal distributions, while CREDENCE produces sharp artificial modes and distorted binary-variable frequencies.

F.4 Constituent validation losses during CAUSALMIX training

Figure 7 and Figure 8 show the constituent validation-loss trajectories for the two CAUSALMIX generators. For the covariate generator, the reconstruction loss decreases rapidly during the initial epochs, while the KL contribution increases initially and then stabilizes; overlap control was not active in this setting. For the potential-outcome generator, the treatment-effect and confounding calibration losses decrease substantially during training, and their residual-variance penalties also decline to low levels. Reconstruction and KL terms stabilize after the initial optimization phase. Although the treatment-effect calibration term exhibits some minibatch-related variability, all causal-control components remain low and stable later in training, with no evidence of divergence among competing objectives. Together, these trajectories indicate that the constituent objectives are jointly optimized and reach stable regimes during training.

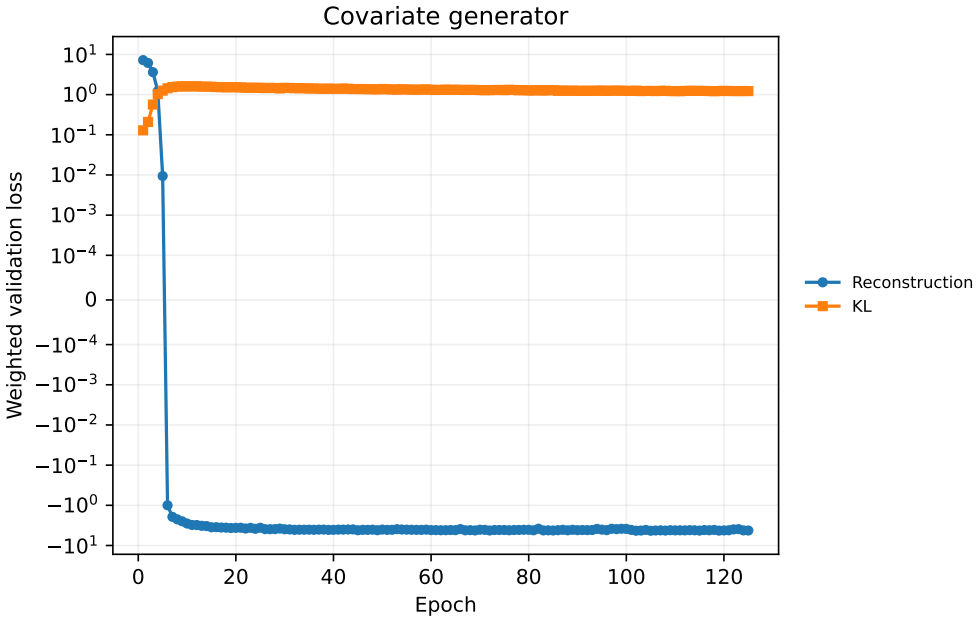


Figure 7: Covariate-generator validation losses during training. Reconstruction and KL contributions are shown over epochs using a symmetric-log scale because the Gaussian reconstruction NLL can take negative values. The reconstruction term decreases rapidly and both components stabilize early in training. Overlap control was not active in this setting.

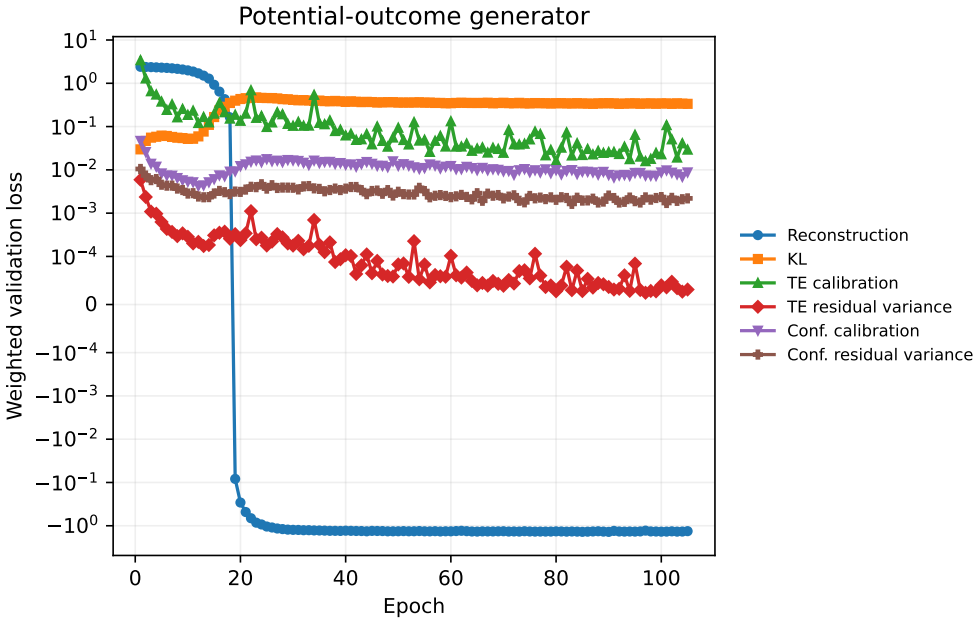


Figure 8: Potential-outcome generator validation losses during training. Weighted reconstruction, KL, treatment-effect calibration and residual-variance, and confounding calibration and residual-variance terms are shown across epochs on a symmetric-log scale. The causal-control losses decrease substantially during early training and subsequently stabilize.

G mCRPC Cohort and Generator Validation Details

G.1 Clinical context

Metastatic castration-resistant prostate cancer (mCRPC) is an advanced stage of prostate cancer characterized by resistance to androgen-deprivation therapy and is associated with rapid disease progression, poor prognosis, and limited overall survival [37]. The most commonly used first-line treatments for mCRPC are the androgen receptor–targeted therapies abiraterone and enzalutamide, which are frequently prescribed interchangeably in routine practice due to the absence of head-to-head phase III randomized trials. Both therapies are associated with clinically meaningful adverse events, and emerging observational evidence suggests that their safety profiles may differ across patient subgroups [38].

G.2 Data description

Empirical data were obtained from the Optum Clinformatics Data Mart, a longitudinal administrative claims database containing member enrollment, demographics, inpatient and outpatient claims, pharmacy claims, laboratory results, and provider information. Claims from January 1, 2014 to March 31, 2024 were used.

Patients with mCRPC were identified using a validated claims-based algorithm proposed by Freedland et al. [39], updated to include recently approved therapies. The analytic cohort comprised 4,098 patients initiating first-line treatment with either enzalutamide or abiraterone after the mCRPC index date.

The outcome was treatment-related hospitalization or emergency department use during follow-up. Treatment was defined by first filled prescription after the index date: abiraterone initiators were treated and enzalutamide initiators were controls. Baseline covariates included 18 variables spanning binary, categorical, and continuous types, including age, race, metastatic site, prior prostate cancer treatments, prior abiraterone or enzalutamide use, Charlson comorbidity index, cardiovascular disease, diabetes, other clinical conditions, and concomitant medications.

G.3 Training details

The CAUSALMIX generator used one hidden layer with 64 units in both encoder and decoder, batch size 10, and latent dimension equal to the number of target variables. Treatment-effect and confounding rigidity parameters were set around 10^3, while overlap rigidity was set between 10^1 and 10^2. Treatment-effect and confounding losses used MSE, Smooth L1, and residual-variance regularization. Training used early stopping on validation loss, and robustness was assessed through validation-loss convergence and consistency of causal quantities across generated datasets.

G.4 Causal scenario definitions

Scenario 1. Homogeneous treatment effect, no confounding, and perfect overlap:

$$\tau(X) \equiv 0.1, \qquad \kappa(X,T) \equiv 0, \qquad \log \alpha(X) \equiv 0.$$

Scenario 2. Linear heterogeneous effect, mild confounding, and moderate constant overlap:

$$\tau(X) = 0.05 + 0.015\,\mathrm{cvd} + 0.01\,\mathrm{age} + 0.01\,\mathrm{Charlson}, \qquad \kappa(X,T) \equiv 0.02, \qquad \log \alpha(X) \equiv 1.$$

Scenario 3. Nonlinear heterogeneous effect, covariate-dependent confounding, and covariate-dependent overlap:

$$\tau(X) = 0.02 + 0.05 \tanh(0.4\,\mathrm{cvd} + 0.2\,\mathrm{age} + 0.2\,\mathrm{Charlson} + 0.4\,\mathrm{dementia}),$$
$$\kappa(X,T) = 0.05 - 0.01(2T-1)\tanh(0.5\,\mathrm{Charlson} + 0.6\,\mathrm{cvd} + 0.2\,\mathrm{age}),$$
$$\log \alpha(X) = 2(2\,\mathrm{Abiraterone_prev} - 1).$$

G.5 Distributional fidelity results

Tables 6–8 report distributional fidelity across the three validation scenarios.

Table 6: Distributional fidelity using BGMM and standard Gaussian priors in Scenario 1.

Category	Metric	Direction	BGMM	Gaussian
Marginal (cont.)	Normalized Wasserstein (mean)	↓ better	0.071 (0.001)	0.336 (0.010)
Marginal (cont.)	KSComplement (mean)	↑ better	0.939 (0.001)	0.853 (0.006)
Marginal (disc.)	TVComplement (mean)	↑ better	0.973 (0.001)	0.939 (0.001)
Pairwise (cont–cont.)	CorrelationSimilarity	↑ better	0.987 (0.002)	0.915 (0.005)
Pairwise (all vars)	SU similarity (mean)	↑ better	0.994 (0.002)	0.992 (0.000)
Pairwise (disc–disc.)	ContingencySimilarity (mean)	↑ better	0.953 (0.001)	0.899 (0.002)
Conditional (all except C)[a]	Weighted MMD^2	↓ better	0.004 (0.001)	0.007 (0.002)
Conditional (all except C)[a]	Normalized MMD^2 ratio vs real	↓ better	1.533 (0.154)	2.531 (0.601)
Joint (all vars)	Normalized Energy Distance	↓ better	0.004 (0.001)	0.025 (0.001)
Joint (all vars)	C2ST (AUC complement)	↑ better	0.748 (0.009)	0.551 (0.013)

[a] Conditional variable (C) is exposure.

Table 7: Distributional fidelity using BGMM and standard Gaussian priors in Scenario 2.

Category	Metric	Direction	BGMM	Gaussian
Marginal (cont.)	Normalized Wasserstein (mean)	↓ better	0.111 (0.004)	0.360 (0.011)
Marginal (cont.)	KSComplement (mean)	↑ better	0.957 (0.001)	0.858 (0.004)
Marginal (disc.)	TVComplement (mean)	↑ better	0.978 (0.001)	0.928 (0.002)
Pairwise (cont–cont.)	CorrelationSimilarity	↑ better	0.985 (0.004)	0.950 (0.007)
Pairwise (all vars)	SU similarity (mean)	↑ better	0.993 (0.002)	0.990 (0.000)
Pairwise (disc–disc.)	ContingencySimilarity (mean)	↑ better	0.956 (0.001)	0.881 (0.003)
Conditional (all except C)[a]	Weighted MMD^2	↓ better	0.006 (0.001)	0.016 (0.001)
Conditional (all except C)[a]	Normalized MMD^2 ratio vs real	↓ better	2.244 (0.446)	5.981 (0.358)
Joint (all vars)	Normalized Energy Distance	↓ better	0.005 (0.000)	0.033 (0.002)
Joint (all vars)	C2ST (AUC complement)	↑ better	0.699 (0.016)	0.510 (0.016)

[a] Conditional variable (C) is exposure.

Table 8: Distributional fidelity using BGMM and standard Gaussian priors in Scenario 3.

Category	Metric	Direction	BGMM	Gaussian
Marginal (cont.)	Normalized Wasserstein (mean)	↓ better	0.075 (0.001)	0.679 (0.008)
Marginal (cont.)	KSComplement (mean)	↑ better	0.954 (0.001)	0.720 (0.005)
Marginal (disc.)	TVComplement (mean)	↑ better	0.961 (0.001)	0.898 (0.002)
Pairwise (cont–cont.)	CorrelationSimilarity	↑ better	0.984 (0.002)	0.883 (0.009)
Pairwise (all vars)	SU similarity (mean)	↑ better	0.991 (0.002)	0.990 (0.000)
Pairwise (disc–disc.)	ContingencySimilarity (mean)	↑ better	0.928 (0.001)	0.841 (0.003)
Conditional (all except C)[a]	Weighted MMD^2	↓ better	0.023 (0.002)	0.029 (0.002)
Conditional (all except C)[a]	Normalized MMD^2 ratio vs real	↓ better	8.563 (0.604)	10.863 (0.724)
Joint (all vars)	Normalized Energy Distance	↓ better	0.011 (0.000)	0.080 (0.002)
Joint (all vars)	C2ST (AUC complement)	↑ better	0.625 (0.008)	0.333 (0.010)

[a] Conditional variable (C) is exposure.

G.6 Distributional visual diagnostics

Figures 9–12 provide marginal and pairwise distributional diagnostics for Scenario 3.

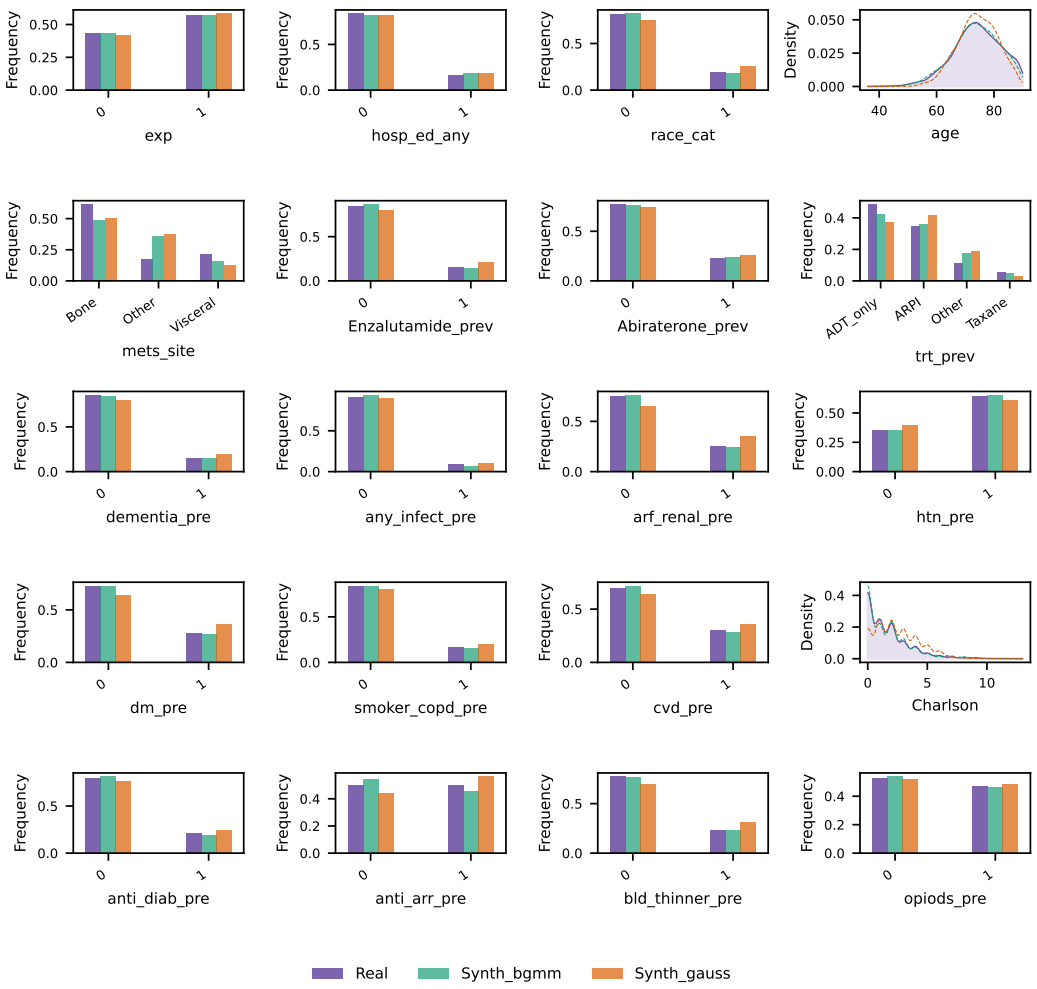


Figure 9: Marginal distributions of all variables in the synthetic data under BGMM and Gaussian priors compared with the real mCRPC dataset (Scenario 3).

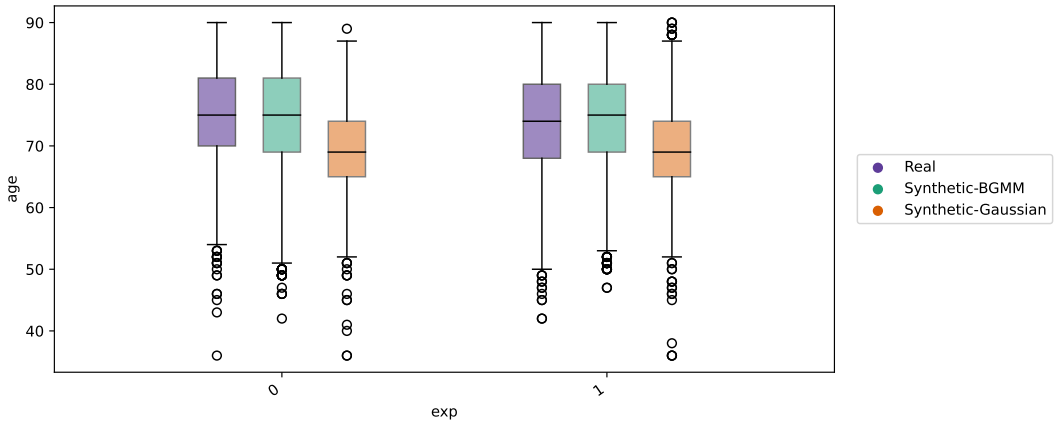


Figure 10: Pairwise distributions of exposure and age in the synthetic data under BGMM and Gaussian priors compared with the real mCRPC dataset (Scenario 3).

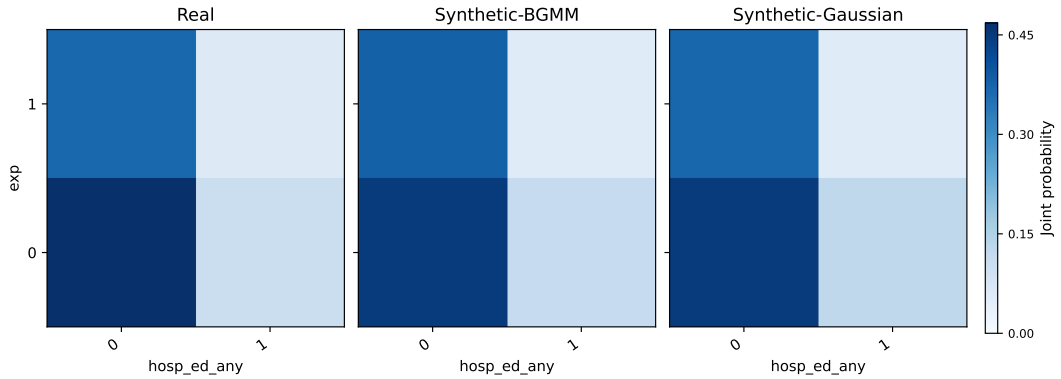


Figure 11: Pairwise distributions of exposure and outcome in the synthetic data under BGMM and Gaussian priors compared with the real mCRPC dataset (Scenario 3).

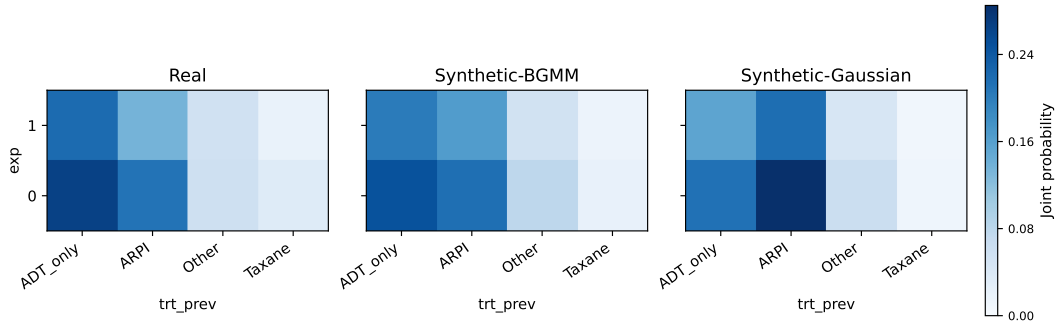


Figure 12: Pairwise distributions of exposure and treatment history in the synthetic data under BGMM and Gaussian priors compared with the real mCRPC dataset (Scenario 3).

G.7 Causal-fidelity results

Tables 9–11 report causal-structure fidelity across validation scenarios. Figures 13 and 14 provide representative confounding and overlap diagnostics.

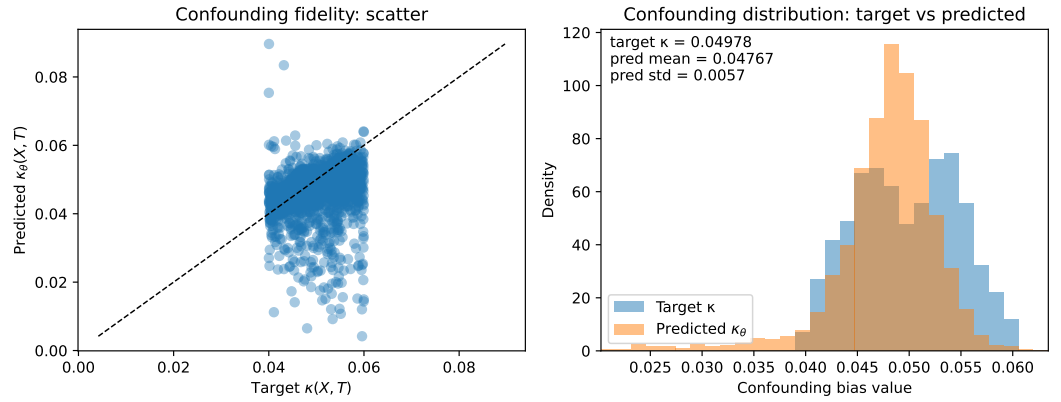


Figure 13: Comparison of generated and target unmeasured confounding of the mCRPC dataset from CAUSALMIX with a BGMM prior in Scenario 3.

Table 9: Causal-structure fidelity of synthetic data using BGMM and standard Gaussian priors in Scenario 1.

Category	Metric	Direction	BGMM	Gaussian
Treatment Effect	CATE/ITE MAE	↓ better	0.003 (0.000)	0.004 (0.000)
Treatment Effect	CATE Correlation	↑ better	NA	NA
Treatment Effect	ATE Error	↓ better	0.000 (0.000)	0.002 (0.000)
Treatment Effect	TE Distribution Distance (W1)	↓ better	0.003 (0.000)	0.004 (0.000)
Confounding	Confounding MAE	↓ better	0.000 (0.000)	0.000 (0.000)
Confounding	Group-wise MAE (T=0)	↓ better	0.000 (0.000)	0.000 (0.000)
Confounding	Group-wise MAE (T=1)	↓ better	0.000 (0.000)	0.000 (0.000)
Confounding	Confounding Dist. (W1)	↓ better	0.000 (0.000)	0.000 (0.000)
Overlap (decoder)	MSE	↓ better	0.000 (0.000)	0.000 (0.000)
Overlap (decoder)	Fraction within tolerance	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Propensity AUC[a]	NA	0.539 (0.007)	0.536 (0.008)
Overlap (propensity)	Histogram overlap coefficient	↑ better	0.940 (0.017)	0.947 (0.015)
Overlap (propensity)	Common support fraction	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=0)	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=1)	↑ better	1.000 (0.000)	1.000 (0.000)

[a] AUC closer to 0.5 indicates stronger overlap.

Table 10: Causal-structure fidelity of synthetic data using BGMM and standard Gaussian priors in Scenario 2.

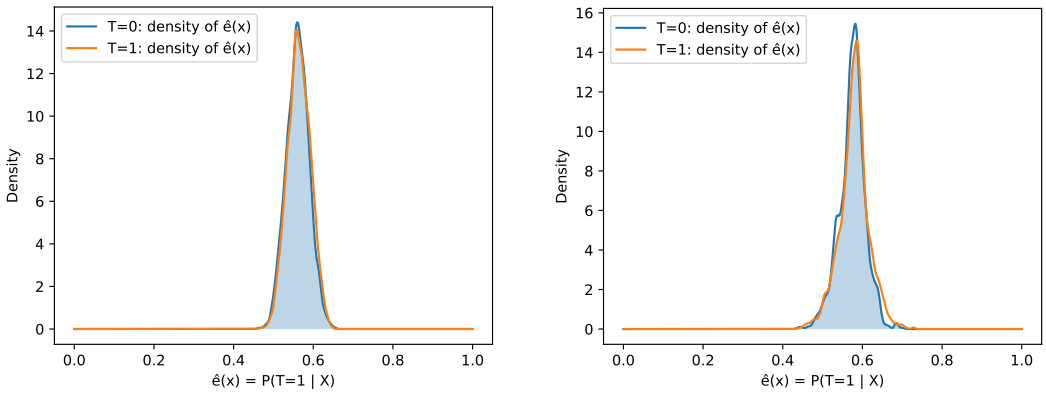
Category	Metric	Direction	BGMM	Gaussian
Treatment Effect	CATE/ITE MAE	↓ better	0.003 (0.000)	0.004 (0.000)
Treatment Effect	CATE Correlation	↑ better	0.971 (0.002)	0.915 (0.004)
Treatment Effect	ATE Error	↓ better	0.000 (0.000)	0.001 (0.000)
Treatment Effect	TE Distribution Distance (W1)	↓ better	0.001 (0.000)	0.001 (0.000)
Confounding	Confounding MAE	↓ better	0.002 (0.000)	0.002 (0.000)
Confounding	Group-wise MAE (T=0)	↓ better	0.002 (0.000)	0.002 (0.000)
Confounding	Group-wise MAE (T=1)	↓ better	0.002 (0.000)	0.002 (0.000)
Confounding	Confounding Dist. (W1)	↓ better	0.002 (0.000)	0.002 (0.000)
Overlap (decoder)	MSE	↓ better	0.004 (0.000)	0.005 (0.000)
Overlap (decoder)	Fraction within tolerance	↑ better	0.997 (0.001)	0.993 (0.001)
Overlap (propensity)	Propensity AUC[a]	NA	0.561 (0.007)	0.563 (0.008)
Overlap (propensity)	Histogram overlap coefficient	↑ better	0.850 (0.021)	0.855 (0.015)
Overlap (propensity)	Common support fraction	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=0)	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=1)	↑ better	1.000 (0.000)	1.000 (0.000)

[a] AUC closer to 0.5 indicates stronger overlap.

Table 11: Causal-structure fidelity of synthetic data using BGMM and standard Gaussian priors in Scenario 3.

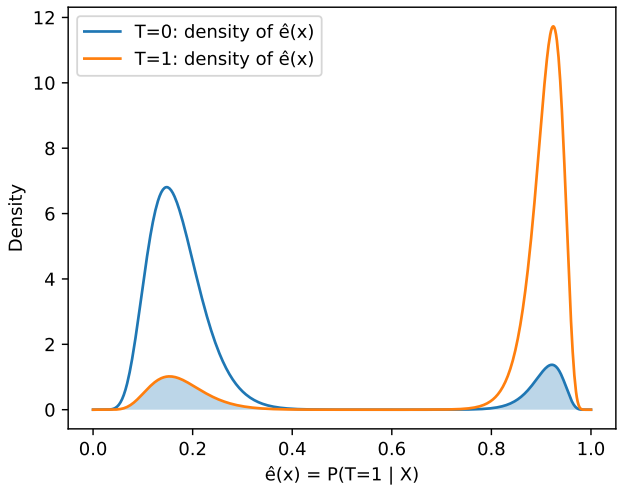
Category	Metric	Direction	BGMM	Gaussian
Treatment Effect	CATE/ITE MAE	↓ better	0.004 (0.000)	0.004 (0.000)
Treatment Effect	CATE Correlation	↑ better	0.956 (0.003)	0.931 (0.005)
Treatment Effect	ATE Error	↓ better	0.001 (0.000)	0.002 (0.000)
Treatment Effect	TE Distribution Distance (W1)	↓ better	0.002 (0.000)	0.002 (0.000)
Confounding	Confounding MAE	↓ better	0.004 (0.000)	0.005 (0.000)
Confounding	Group-wise MAE (T=0)	↓ better	0.005 (0.000)	0.005 (0.000)
Confounding	Group-wise MAE (T=1)	↓ better	0.003 (0.000)	0.004 (0.000)
Confounding	Confounding Dist. (W1)	↓ better	0.002 (0.000)	0.002 (0.000)
Overlap (decoder)	MSE	↓ better	0.004 (0.000)	0.300 (0.091)
Overlap (decoder)	Fraction within tolerance	↑ better	0.996 (0.001)	0.934 (0.004)
Overlap (propensity)	Propensity AUC[a]	NA	0.889 (0.006)	0.859 (0.005)
Overlap (propensity)	Histogram overlap coefficient	↑ better	0.243 (0.012)	0.309 (0.008)
Overlap (propensity)	Common support fraction	↑ better	0.998 (0.003)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=0)	↑ better	1.000 (0.000)	1.000 (0.000)
Overlap (propensity)	Common support fraction (T=1)	↑ better	0.996 (0.004)	1.000 (0.000)

[a] AUC closer to 0.5 indicates stronger overlap.



(a) Scenario 1

(b) Scenario 2



(c) Scenario 3

Figure 14: Generated overlap of the mCRPC datasets from CAUSALMIX with a BGMM prior across causal scenarios. Scenario 1: perfect overlap $\log \alpha(X) = 0$; Scenario 2: moderate constant overlap $\log \alpha(X) = 1$; Scenario 3: covariate-dependent overlap $\log \alpha(X) = 2\big(2\,\mathrm{Abiraterone_prev} - 1\big)$.

G.8 Privacy results

Tables 12–14 report record-level privacy diagnostics across validation scenarios.

Table 12: Privacy of synthetic data using BGMM and standard Gaussian priors in Scenario 1.

Metric	Direction	BGMM	Gaussian
Protection Fraction	↑ better	0.595 (0.009)	0.696 (0.008)
Distance Ratio (mean)	↑ better	2.535 (0.083)	3.254 (0.163)
Distance Ratio (p5)	↑ better	0.233 (0.028)	0.501 (0.003)
Distance Ratio (p50)	↑ better	1.039 (0.006)	1.122 (0.006)
Distance Ratio (p95)	↑ better	4.841 (0.263)	6.150 (0.304)
Standardized Distance Ratio	↑ better	0.276 (0.003)	0.333 (0.003)

Table 13: Privacy of synthetic data using BGMM and standard Gaussian priors in Scenario 2.

Metric	Direction	BGMM	Gaussian
Protection Fraction	↑ better	0.607 (0.007)	0.747 (0.007)
Distance Ratio (mean)	↑ better	2.787 (0.193)	3.723 (0.249)
Distance Ratio (p5)	↑ better	0.255 (0.021)	0.624 (0.036)
Distance Ratio (p50)	↑ better	1.057 (0.007)	1.195 (0.009)
Distance Ratio (p95)	↑ better	4.623 (0.207)	7.681 (0.622)
Standardized Distance Ratio	↑ better	0.282 (0.002)	0.373 (0.007)

Table 14: Privacy of synthetic data using BGMM and standard Gaussian priors in Scenario 3.

Metric	Direction	BGMM	Gaussian
Protection Fraction	↑ better	0.720 (0.006)	0.839 (0.004)
Distance Ratio (mean)	↑ better	3.298 (0.245)	4.287 (0.089)
Distance Ratio (p5)	↑ better	0.396 (0.030)	0.788 (0.014)
Distance Ratio (p50)	↑ better	1.169 (0.007)	1.337 (0.009)
Distance Ratio (p95)	↑ better	6.046 (0.055)	11.740 (0.493)
Standardized Distance Ratio	↑ better	0.297 (0.003)	0.559 (0.005)

H Sensitivity to Causal Rigidity Parameters, Cross-Control Effects, and Feasibility Tests

H.1 One-at-a-time sensitivity to causal rigidity parameters

We evaluate the sensitivity of CAUSALMIX to the three causal rigidity parameters: overlap rigidity λ_α, treatment-effect rigidity λ_τ, and confounding rigidity λ_κ. Under mCRPC Scenario 3, for each parameter, we perform a one-at-a-time sweep while holding the remaining rigidity parameters fixed at their recommended scenario-specific operating values. This design evaluates the realism–control trade-off in the full trained model, rather than in an ablated model where other causal controls are disabled. In all sweeps, the blend weights ($w_\tau = w_\kappa = 0.4$) and variance weights ($\lambda_\tau^{\text{var}} = \lambda_\kappa^{\text{var}} = 0.2$) are held fixed and retain variance regularization.

For each sweep, the selected parameter is varied over a grid including a near-zero value and multiplicative deviations around the default. We compare global distributional fidelity, measured by normalized energy distance, against the corresponding causal calibration error: overlap MSE for λ_α, CATE MAE for λ_τ, and confounding MAE for λ_κ. The default operating value is highlighted in each figure.

Across all three rigidity parameters, the trade-offs are stable and interpretable. Increasing λ_α substantially improves overlap calibration with only modest degradation in distributional realism. Increasing λ_τ improves treatment-effect fidelity while leaving normalized energy distance nearly unchanged. Similarly, increasing λ_κ improves confounding alignment with minimal impact on realism. These results support the use of the rigidity parameters as practical controls for navigating the realism–causal fidelity frontier.

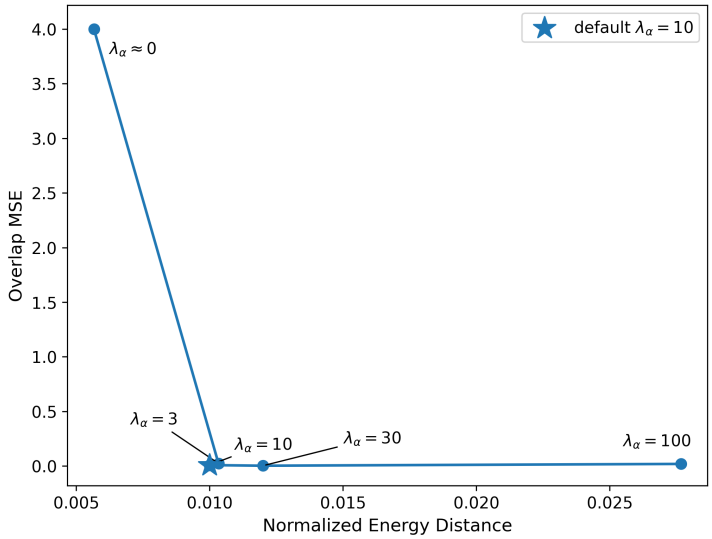


Figure 15: Overlap rigidity sweep. Normalized energy distance measures global distributional fidelity, while overlap MSE measures calibration to the target overlap function. Increasing λ_α substantially improves overlap calibration with only modest loss in realism, revealing a stable operating region around the default setting.

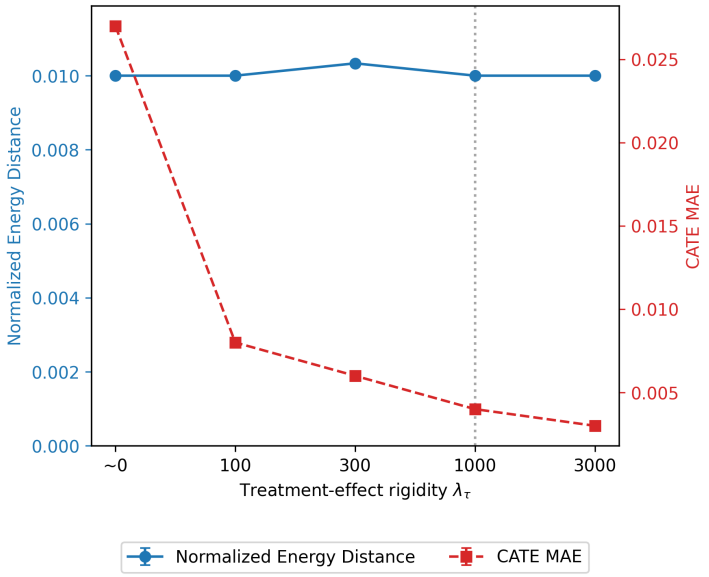


Figure 16: Treatment-effect rigidity sweep. Increasing λ_τ improves CATE fidelity while normalized energy distance remains nearly unchanged, indicating a broad stable operating region for treatment-effect control.

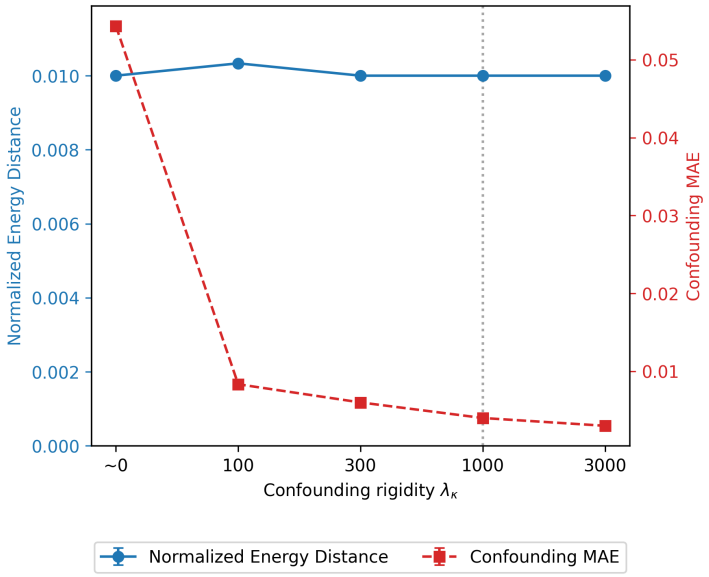


Figure 17: Confounding rigidity sweep. Increasing λ_κ improves alignment with the target confounding function while maintaining nearly constant distributional realism, indicating stable control of unmeasured confounding over the tested range.

H.2 Cross-control sensitivity

We additionally examine interactions among causal controls regarding whether tuning one control affects the others. In nonlinear mCRPC Scenario 3, we vary λ_τ while fixing $\lambda_\alpha = 10$ and $\lambda_\kappa = 1000$, and report treatment-effect calibration together with distributional realism, confounding and overlap calibration (Table 15).

Table 15: Cross-control sensitivity in mCRPC Scenario 3 with the BGMM prior. Values are mean (sd) over generated datasets. Arrows indicate preferred direction.

λ_τ	Energy dist. ↓	CATE MAE ↓	CATE corr. ↑	Conf. MAE ↓	Overlap MSE ↓
0	0.0114 (0.0022)	0.0281 (0.0012)	0.416 (0.111)	0.0063 (0.0006)	0.0058 (0.0023)
100	0.0113 (0.0020)	0.0079 (0.0002)	0.841 (0.020)	0.0030 (0.0000)	0.0058 (0.0029)
300	0.0114 (0.0025)	0.0057 (0.0006)	0.907 (0.017)	0.0030 (0.0000)	0.0060 (0.0026)
1000	0.0117 (0.0023)	0.0040 (0.0000)	0.952 (0.005)	0.0037 (0.0006)	0.0056 (0.0022)
3000	0.0114 (0.0019)	0.0030 (0.0000)	0.973 (0.006)	0.0046 (0.0004)	0.0056 (0.0022)

As λ_τ increases from approximately zero to 3000, CATE MAE decreases from 0.0281 to 0.0030 and CATE correlation increases from 0.416 to 0.973, so λ_τ primarily controls the intended treatment-effect property. Normalized energy distance remains near 0.011, overlap MSE near 0.006, and confounding MAE below 0.007 throughout. When the requested target is compatible with the data distribution, stronger control is therefore not accompanied by a loss of realism, and the other causal errors stay small. A trade-off emerges only when the requested properties are difficult to reconcile with the empirical distribution, as examined next.

H.3 Feasibility-boundary stress tests

Joint feasibility of the requested causal targets cannot be guaranteed a priori: the objective returns the closest feasible generator under the chosen weights, and the cost of an infeasible request can fall on realism, on another causal property, or on both. We illustrate this behavior with two stress tests.

Test 1: competing controls in mCRPC Scenario 3. We impose a near-perfect overlap target ($\log \alpha(X) = 0$) together with a fourfold stronger confounding target, fix $\lambda_\tau = 1000$, and increase λ_α and λ_κ together (Table 16). This requires the generator to preserve the empirical mixed-type joint distribution, make the treated and control covariate distributions nearly indistinguishable, and impose a strong treatment-arm-specific counterfactual contrast.

Table 16: Feasibility-boundary stress test in mCRPC Scenario 3 with the BGMM prior, using a near-perfect overlap target ($\log \alpha(X) = 0$) and a fourfold stronger confounding target. $\lambda_\tau = 1000$ is fixed. Values are mean (sd) over generated datasets. Arrows indicate preferred direction.

Setting	λ_α	λ_κ	Energy dist. ↓	Overlap MSE ↓	Conf. MAE ↓	CATE MAE ↓
None	0	0	0.0036 (0.0011)	0.0978 (0.0255)	0.2051 (0.0007)	0.0030 (0.0002)
Weak	3	100	0.0036 (0.0005)	0.0003 (0.0005)	0.0208 (0.0017)	0.0039 (0.0004)
Default	10	1000	0.0038 (0.0007)	0.0000 (0.0000)	0.0108 (0.0007)	0.0084 (0.0005)
Strong	30	3000	0.0042 (0.0004)	0.0000 (0.0000)	0.0060 (0.0002)	0.0129 (0.0006)
Extreme	100	10000	0.0052 (0.0007)	0.0000 (0.0000)	0.0028 (0.0002)	0.0156 (0.0003)

Increasing rigidity drives overlap and confounding toward their targets: confounding MAE falls from 0.2051 to 0.0028, and overlap MSE falls from 0.0978 to approximately zero. The cost appears in the other objectives: CATE MAE rises from 0.0030 to 0.0156 and energy distance rises more modestly from 0.0036 to 0.0052. This illustrates the closest-feasible-generator behavior: as the requested controls become more stringent, CAUSALMIX moves closer to the requested regime, but the cost may shift to another causal property or to realism.

Test 2: incompatible treatment-effect target in a simulation with known truth. To examine incompatible targets more directly, we use a simulated dataset with a transparent, known data-generating process. We generate $n = 5{,}000$ observations with mixed continuous and binary covariates,

moderate treatment overlap, a nonlinear baseline outcome model, no unmeasured confounding, and a true heterogeneous treatment effect $\tau(X) = 1 + 0.5X_1$. In all settings, we retain the correct confounding and overlap targets and hold the rigidity parameters fixed at $\lambda_\tau = 1000$, $\lambda_\kappa = 1000$, and $\lambda_\alpha = 100$. We vary only the requested treatment-effect function, moving from the correct target to increasingly incompatible mild, moderate, and extreme targets, and assess both target attainment and divergence from the reference distribution (Table 17).

Table 17: Feasibility-boundary stress test in a simulation with a known data-generating process ($n = 5{,}000$; true effect $\tau(X) = 1 + 0.5X_1$). Confounding and overlap targets are correct in all rows, and $\lambda_\tau = 1000$, $\lambda_\kappa = 1000$, $\lambda_\alpha = 100$ are fixed. Arrows indicate preferred direction.

Setting	Requested $\tau(X)$	CATE MAE ↓	Conf. MAE ↓	Overlap MSE ↓	Energy dist. ↓	$W_1, Y \mid T = 0$ ↓	$W_1, Y \mid T = 1$ ↓
Correct	$1 + 0.5X_1$	0.0181	0.0021	0.0030	0.0166	0.5658	0.1327
Mild violation	$1.5 + 0.75X_1$	0.0303	0.0043	0.0030	0.0280	0.5830	0.3385
Moderate violation	$2.5 + 1.5X_1$	0.0370	0.0039	0.0030	0.0255	0.7465	0.3825
Extreme violation	$5 - 3X_1$	0.0422	0.0067	0.0030	0.0497	0.1418	1.3867

W_1: 1-Wasserstein distance between observed and generated outcomes within each treatment arm.

CAUSALMIX continues to approximate the requested target in all settings, although CATE MAE (measured against the requested target) increases from 0.018 for the correct target to 0.042 for the extreme target. The other explicitly controlled properties remain stable, with confounding MAE below 0.007 and overlap MSE approximately 0.003 throughout. However, the most incompatible target is accompanied by a substantial loss of distributional fidelity: energy distance increases from 0.017 to 0.050, and the treated-arm outcome Wasserstein distance increases from 0.133 to 1.387. When a strongly enforced causal target conflicts with the reference distribution, the model may approach that target while preserving the other controlled causal properties at the cost of reduced observed-data fidelity.

Summary. Together, these tests show that CAUSALMIX can accommodate a broad range of target combinations, but that joint feasibility is revealed through the multi-objective frontier rather than guaranteed in advance. We therefore recommend evaluating target-attainment and fidelity diagnostics together, rather than assuming that all requested regimes are simultaneously achievable.

I CATE Estimator Benchmark Details

I.1 Estimator families

We evaluate ten CATE estimators spanning four families: meta-learners, doubly robust learners, double machine learning estimators, and tree-based methods. X-learners use lasso and gradient boosting regression (GBR) final-stage CATE models. DML and DR learners use linear, lasso, and GBR final-stage CATE models. Tree-based estimators include causal forests and Bayesian causal forests (BCF).

For all non-tree estimators, gradient boosting trees are used to estimate nuisance functions, including propensity scores and outcome regressions. Cross-fitting is applied throughout.

I.2 Preprocessing

We construct two feature representations for each synthetic dataset. Dummy encoding is used for X-learner, DML, and DR variants to avoid multicollinearity in linear final-stage models. Full one-hot encoding is used for causal forests and BCF. Continuous covariates are standardized and binary variables are left unchanged.

I.3 Inference and metrics

Asymptotic analytic inference is used for linear final-stage CATE models. Bootstrap inference with 100 resamples is used for lasso and nonparametric GBR final-stage models. Each scenario is evaluated over 50 independent synthetic replications. All estimators are run single-threaded and wall-clock runtime is recorded.

For ATE estimation, we report ATE RMSE, mean bias, estimated standard error, and confidence interval coverage. For CATE estimation, we report PEHE,

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) - \tau(X_i)\right)^2},$$

pointwise CATE coverage, and runtime.

I.4 Estimator benchmark results

Tables 18 and 19 report estimator accuracy and coverage under no unmeasured confounding and moderate unmeasured confounding. Figures 18 and 19 show ATE bias and PEHE across estimators.

Table 18: Accuracy and coverage of causal estimators under no unmeasured confounding.

Estimator	Runtime (s , SD)	ATE RMSE	Mean ATE SE	ATE Coverage	Mean CATE Coverage
Bayesian Causal Forest	258.7 (9.5)	0.010	0.010	0.90	0.91
Causal Forest	25.1 (1.7)	0.009	0.009	0.94	0.87
Standard Linear DML	2.2 (0.1)	0.009	0.009	0.96	0.90
Lasso DML	229.9 (10.1)	0.009	0.008	0.88	0.37
GBR DML	268.1 (10.7)	0.010	0.113	1.00	0.82
Standard Linear DR-Learner	2.3 (0.1)	0.010	0.046	1.00	0.92
Lasso DR-Learner	234.5 (10.6)	0.010	0.009	0.88	0.40
GBR DR-Learner	274.0 (10.8)	0.010	0.125	1.00	0.82
Lasso X-Learner	118.4 (4.1)	0.024	0.036	1.00	0.82
GBR X-Learner	160.4 (4.4)	0.024	0.085	1.00	0.82

Table 19: Accuracy and coverage of causal estimators under moderate unmeasured confounding.

Estimator	Runtime (s , SD)	ATE RMSE	Mean ATE SE	ATE Coverage	Mean CATE Coverage
Bayesian Causal Forest	260.6 (9.6)	0.023	0.011	0.48	0.86
Causal Forest	25.3 (1.8)	0.027	0.009	0.18	0.80
Standard Linear DML	2.1 (0.1)	0.026	0.009	0.24	0.84
Lasso DML	225.0 (7.9)	0.025	0.010	0.28	0.26
GBR DML	262.8 (7.7)	0.026	0.125	1.00	0.83
Standard Linear DR-Learner	2.2 (0.1)	0.028	0.046	1.00	0.85
Lasso DR-Learner	230.5 (8.4)	0.028	0.008	0.14	0.25
GBR DR-Learner	269.7 (7.3)	0.028	0.132	1.00	0.81
Lasso X-Learner	116.4 (3.8)	0.008	0.035	1.00	0.86
GBR X-Learner	158.6 (4.3)	0.009	0.088	1.00	0.83

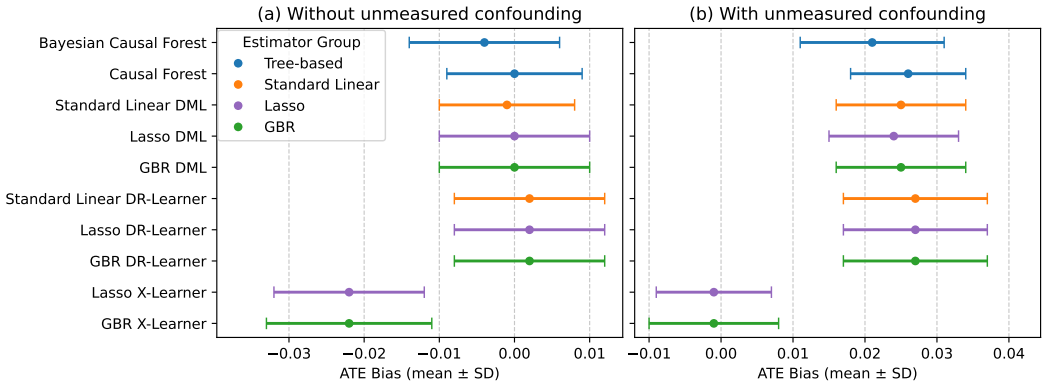


Figure 18: Comparison of ATE bias across estimators under no unmeasured confounding ($\kappa = 0$) and moderate unmeasured confounding ($\kappa = 0.02$). Points denote mean bias across 50 independent replications and error bars indicate one standard deviation. Estimator groups are defined based on the CATE model.

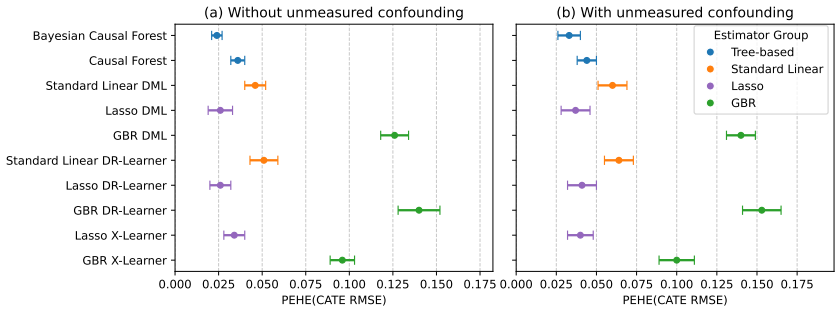


Figure 19: Comparison of PEHE across estimators under no unmeasured confounding ($\kappa = 0$) and moderate unmeasured confounding ($\kappa = 0.02$). Points denote mean RMSE across 50 independent replications and error bars indicate one standard deviation. Estimator groups are defined based on the CATE model.

J Additional Downstream Applications

J.1 Hyperparameter tuning

Beyond estimator comparison, CAUSALMIX can tune causal-estimator hyperparameters by evaluating directly against known ground truth under an empirically calibrated DGP. We use the calibrated no-confounding mCRPC generator from the CATE benchmarking experiment, with linear treatment-effect heterogeneity and moderate overlap, as a surrogate ground-truth environment for tuning causal forests.

Design. We vary two causal-forest hyperparameters that govern smoothness, stability, and computational cost:

$$\texttt{min.node.size} \in \{5, 10, 20, 40, 80\}, \qquad \texttt{num.trees} \in \{500, 1000, 2000, 4000\}.$$

For each configuration, we generate 50 independent synthetic datasets from the calibrated generator, fit a causal forest, and evaluate ATE RMSE and coverage, PEHE (CATE RMSE), mean absolute CATE bias, mean CATE confidence interval width, CATE coverage, and runtime. Hyperparameter selection prioritizes CATE accuracy and calibration, while ATE metrics serve as population-level diagnostics.

Results. ATE estimation is largely insensitive to the hyperparameter grid: ATE RMSE remains approximately 0.009 and ATE coverage stays within a narrow range of 0.92–0.94 across configurations. In contrast, CATE performance is primarily driven by the minimum leaf size. Increasing `min.node.size` improves CATE point accuracy, with PEHE decreasing from approximately 0.036–0.037 at leaf size 5 to approximately 0.023 at leaf size 80 (Figure 20). However, this smoother estimation comes at the cost of reduced CATE interval coverage, which decreases from approximately 0.87–0.91 to approximately 0.75–0.76.

The same pattern is reflected in Figure 21: larger leaves reduce both mean absolute CATE bias and CATE interval width, suggesting that small leaves produce noisier and more variable heterogeneous-effect estimates in this setting. Once the number of trees is moderately large, `num.trees` has little effect on ATE or CATE accuracy, but it increases runtime roughly linearly: about 7 seconds at 500 trees, 11 seconds at 1,000 trees, 21 seconds at 2,000 trees, and 42 seconds at 4,000 trees.

Takeaway. For heterogeneous safety-effect estimation in this calibrated mCRPC setting, tuning `min.node.size` is more consequential than substantially increasing `num.trees`. A practical operating region is around 1,000–2,000 trees with moderate leaf sizes (e.g., 20–40), which balances CATE accuracy, uncertainty calibration, and runtime.

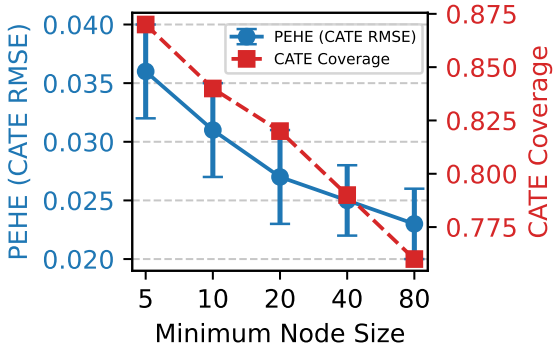


Figure 20: PEHE and CATE coverage versus minimum leaf size for causal forests with 2,000 trees. Error bars indicate ± 1 standard deviation across 50 synthetic replications.

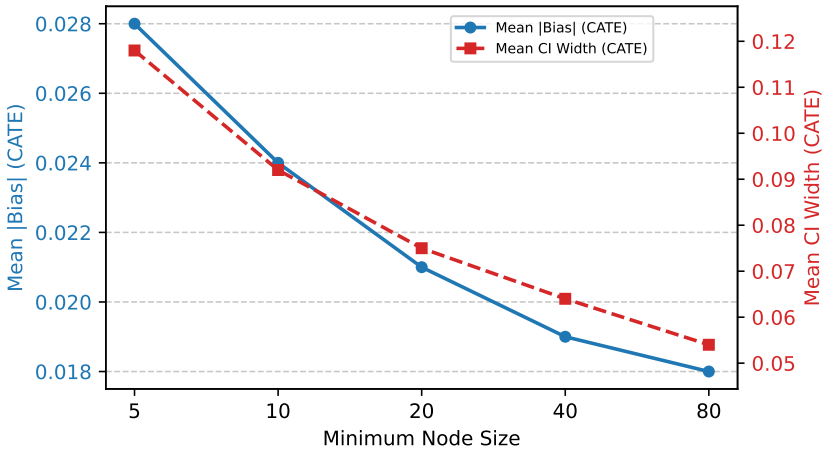


Figure 21: Sensitivity of CATE accuracy and uncertainty to the minimum leaf size in causal forests with 2,000 trees. The solid curve shows mean absolute CATE bias and the dashed curve shows mean CATE confidence interval width, averaged over 50 synthetic replications.

J.2 Simulation-based power analysis

We also use the calibrated generator for prospective study design. The goal is to estimate the sample size required to detect treatment-effect modification in a realistic observational cohort while preserving the empirical covariate distribution, treatment assignment structure, and outcome variability [40].

Target estimand. We focus on baseline cardiovascular disease (CVD) as a pre-specified binary effect modifier. The target estimand is the subgroup-average CATE contrast

$$\Delta = \mathbb{E}[\tau(X) \mid \text{CVD} = 1] - \mathbb{E}[\tau(X) \mid \text{CVD} = 0].$$

For clarity, this study-design experiment uses the no-confounding calibrated regime. The same framework can also vary heterogeneity strength by scaling the centered CATE component

$$h(x) = \tau(x) - \mathbb{E}[\tau(X)]$$

via

$$\tau_\eta(x) = \mathbb{E}[\tau(X)] + \eta h(x),$$

although the results below use the default setting $\eta = 1$.

Design. We consider sample sizes

$$n \in \{200, 500, 1000, 2000, 5000\}.$$

For each n, we generate 100 independent synthetic datasets from the calibrated CausalMix generator. In each replication, we fit a causal forest with 2,000 trees and minimum node size 80. This configuration was chosen to reduce CATE noise and stabilize effect-modifier signals rather than to optimize interval calibration.

Effect modification is tested using the best linear projection procedure implemented in `grf`. Specifically, estimated CATEs are projected onto baseline covariates using doubly robust scores, and inference focuses on the coefficient for `CVD`. Power is estimated as the empirical rejection frequency of the two-sided Wald test at level $\alpha = 0.05$. We also record the variable-importance rank of `CVD` in each replication to assess whether the fitted forest reliably prioritizes the true modifier.

Results. Power increases sharply with sample size (Figure 22). It is low at $n = 200$ (0.13), moderate at $n = 500$ (0.54) and $n = 1000$ (0.67), and high by $n = 2000$ (0.94), approaching 1.0 at $n = 5000$. Variable-importance ranks stabilize more slowly: `CVD` is not reliably prioritized when $n \leq 1000$ (mean rank about 15), improves by $n = 2000$ (mean rank about 8), and is consistently identified as the top modifier by $n = 5000$.

Takeaway. In this calibrated mCRPC setting, detecting the presence of effect modification requires fewer samples than reliably identifying and prioritizing the correct modifier. Approximately 2,000 patients are sufficient for high power to detect CVD effect modification, while stable modifier prioritization requires larger cohorts, closer to 5,000 patients. This illustrates how a validated causal generator can support prospective study design beyond estimator benchmarking.

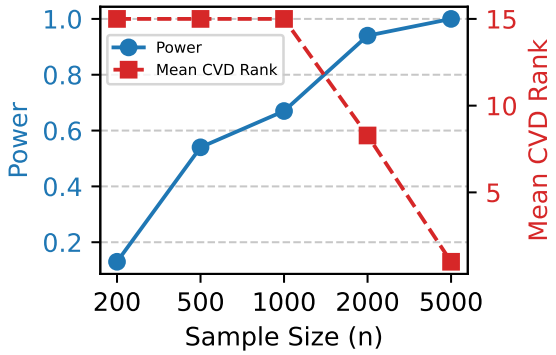


Figure 22: Power to detect CVD effect modification and mean CVD variable-importance rank as sample size increases. Power is estimated over 100 synthetic replications at each sample size.

K Compute Resources

All mCRPC generator-training experiments were run on a single NVIDIA L40S GPU with 64GB system memory. A single mCRPC CAUSALMIX generator training run required approximately 15 minutes, corresponding to about 0.25 GPU-hours. Runtime scaled approximately linearly with the number of causal scenarios, random seeds, and rigidity-parameter settings.

The public benchmark comparison and rigidity-parameter sensitivity analyses were run on a single NVIDIA L40S GPU with 128GB system memory. A single public-benchmark CAUSALMIX fitting run required approximately 4 minutes.

Downstream CATE-estimator experiments were run on CPU. Estimator-level runtimes are reported in Tables 18 and 19.

NeurIPS Paper Checklist

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 state that the contribution is a controllable generative sandbox for mixed-type tabular causal data, with explicit controls for overlap, treatment effects, and unmeasured confounding. The claims are supported by the public benchmark comparison, mCRPC generator validation, and applications of CATE-estimator benchmarking, estimator tuning and power analysis in Sections 4 and Appendices F–J.

2. **Limitations**

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

   Answer: [Yes].

   Justification: The Discussion and Conclusion section explicitly discusses limitations, including reliance on user-specified causal functions, possible incompatibility between causal targets and empirical distributions, limited preservation of rare or high-order structure, and the need to retrain for new causal regimes.

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: [N/A].

   Justification: The paper does not present new formal theorems or proof-based theoretical results. The mathematical content defines the generative objective, causal-control functions, and evaluation diagnostics rather than proving theoretical guarantees.

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: The paper describes the architecture, objective, causal targets, training procedure, benchmark construction, evaluation metrics, hyperparameters, and estimator settings in the main text and appendices. Details are provided in Sections 3–4 and Appendices B–J.

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: Code for the public LaLonde benchmark experiments and the CAUSALMIX framework is available at the repository linked in the Introduction. The LaLonde/NSW–PSID benchmark is publicly accessible, and the paper provides detailed model, training, and evaluation specifications. The mCRPC claims data are licensed and cannot be redistributed; we describe the cohort, variables, and causal scenarios in Appendix G.

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: Training details, optimizer, train–validation split, latent priors, rigidity parameters, causal scenario definitions, estimator settings, and evaluation metrics are described in the main text and Appendices C, F, G, H and I.

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: Tables report mean and standard deviation across repeated generated datasets or replications. Estimator benchmarking and downstream application figures report standard-deviation error bars where applicable; rigidity-parameter sweeps are reported as diagnostic trade-off curves.

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: Appendix K reports the GPU type, memory, per-run generator training time, and public benchmark runtime. Estimator-level CPU runtimes are reported in Table 18 and Table 19.

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: The work is methodological and uses public benchmark data plus de-identified, access-controlled administrative claims data. The paper discusses limitations, privacy diagnostics, and appropriate interpretation of synthetic causal data, and does not release identifiable patient-level data.

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: The paper discusses positive impacts for causal estimator benchmarking, sensitivity analysis, and study design, especially in settings where counterfactual ground truth is unavailable. It also discusses risks, including incorrect user-specified causal assumptions, misleading conclusions if synthetic regimes are treated as truth, and privacy concerns addressed through DCR diagnostics.

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: The paper does not release high-risk pretrained models, image generators, scraped datasets, or identifiable clinical data. The restricted clinical data are not redistributed, and synthetic data quality is evaluated with privacy diagnostics.

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: [Yes].

    Justification: The public benchmark sources are cited, including LaLonde and Dehejia–Wahba. The mCRPC case study uses licensed administrative claims data under applicable data-use restrictions. Python dependencies are listed in the repository's requirements file.

13. **New assets**

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

Answer: [Yes].

Justification: The paper introduces the CAUSALMIX framework, released under the Apache-2.0 license at the repository linked in the Introduction. The repository includes a README with installation instructions, usage examples, and scripts for the reported experiments, and the paper documents the model architecture, objectives, training procedure, evaluation metrics, and experimental settings in Sections 3–4 and Appendices A–I. No new datasets are released; the mCRPC claims data are licensed and not redistributed, and no synthetic patient-level data derived from them are released.

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: The paper does not involve crowdsourcing, participant-facing experiments, or new data collection from human subjects. The clinical case study uses secondary, de-identified administrative claims data.

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: The study does not recruit participants or collect new human-subject data. The clinical component uses de-identified secondary administrative claims data under applicable data-use restrictions; no identifiable patient data are released.

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: LLMs are not a component of the core method, experiments, or evaluation. As a matter of good practice we disclose the following. Generative AI tools (including DeepSeek, ChatGPT and Claude) were used to assist with language editing, manuscript refinement, code development, and discussions of methodological concepts. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript.