Causal-ICM: A Data Fusion Framework For Heterogeneous Treatment Effect Estimation With Multi-Task Gaussian Processes

Evangelos Dimitriou EVANGELOS.DIMITRIOU.22@UCL.AC.UK
Department of Statistical Science, University College London

Edwin Fong CHEFONG@HKU.HK
Department of Statistics and Actuarial Science, University of Hong Kong

Jens Magelund Tarp JQMT@NOVONORDISK.COM
Novo Nordisk A/S

Karla Diaz-Ordaz KARLA.DIAZ-ORDAZ@UCL.AC.UK
Department of Statistical Science, University College London

Brieuc Lehmann B.LEHMANN@UCL.AC.UK
Department of Statistical Science, University College London

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Bridging the gap between internal and external validity is crucial for heterogeneous treatment effect estimation. Randomised controlled trials (RCTs), favoured for their internal validity due to randomisation, often encounter challenges in generalising findings due to strict eligibility criteria. Observational studies, on the other hand, may provide stronger external validity through larger and more representative samples but can suffer from compromised internal validity due to unmeasured confounding. Motivated by these complementary characteristics, we propose a novel Bayesian nonparametric approach, *Causal-ICM*, leveraging multi-task Gaussian processes to integrate data from both RCTs and observational studies. In particular, we introduce a parameter that controls the degree of borrowing between the datasets and prevents the observational dataset from dominating the estimation. We propose a data-adaptive procedure for choosing the optimal value of the parameter. *Causal-ICM* outperforms other data fusion methods in point estimation across the covariate support of the observational study and provides principled uncertainty quantification for the estimated treatment effects. We demonstrate the robust performance of *Causal-ICM* in diverse scenarios through multiple simulation studies and a real-world study.

Keywords: Data Fusion, Bayesian Nonparametrics, Multitask Gaussian Processes, Generalisability, Heterogeneous Treatment Effects

1. Introduction

Treatment effect estimation is an important task in many applications, including medicine, epidemiology, and the social sciences. The goal is to quantify the expected impact of a particular intervention in a given target population. The effect of an intervention within a population may vary systematically with respect to a particular set of covariates. Understanding such treatment effect heterogeneity is critical to ensure that appropriate decisions are taken for all individuals, not just those similar to the average (Brantner et al., 2023). An appropriate characterisation of the uncertainty of treatment effect estimates is also required

to support robust and reliable decision-making. Uncertainty quantification for heterogeneous treatment effects is challenging, however, especially when making inferences about individuals not represented in the study. Classical methods typically rely heavily on extrapolation without adequate variance inflation outside the study support (Degtiar and Rose, 2023).

The gold standard for treatment effect estimation remains the RCT. The random treatment allocation in the study guards against the effect of confounding As a result, under standard identifiability assumptions, we can obtain an unbiased estimate of the treatment effect. However, RCTs have important drawbacks (Frieden, 2017), including high financial costs or the limited sample sizes. While often powered to detect the Average Treatment Effect (ATE), they are typically underpowered for subgroup effects. Furthermore, strict eligibility criteria and selection bias can limit representativeness of the target population (National Academies of Sciences, Engineering, and Medicine, 2022).

Observational studies offer an alternative source of data for treatment effect estimation, typically benefiting from larger sample sizes and better representativeness of the target population. However, they may be susceptible to unobserved confounding - unmeasured variables that affect both outcomes and treatment assignment - leading to biased treatment effect estimates. For example, a patient's underlying health status might influence both the treatment they receive and their eventual outcome, yet remains unmeasured.

The complementary nature of RCT and observational data indicates the potential benefits from combining the two sources to obtain better heterogeneous treatment effect estimates. Such data fusion approaches have received increased interest in recent years. A range of methods have been developed to integrate information from multiple sources of data, including a limited number targeting causal inference problems (Colnet et al. (2024), Lin et al. (2024)). These typically rely on strong, untestable assumptions about the structure of the confounding effect and, moreover, are often not able to provide uncertainty quantification around the point estimates of heterogeneous treatment effects.

In this paper, we introduce a Bayesian nonparametric approach to obtain both point and uncertainty estimates of heterogeneous treatment effects for the target population. Our approach is based on a multi-output Gaussian process (GP), a natural choice in the data fusion context as it enables joint modelling of the RCT and observational outcome regression functions, and uses the Bayesian machinery to share information and quantify uncertainty of the functions in regions of differing covariate support. Our key contributions are as follows:

- We propose *Causal-ICM*, a multi-task GP model for heterogeneous treatment effect estimation that accurately captures complex functional relationships in the presence of unobserved confounding both within the support of the RCT covariate distribution and when extrapolating beyond it.

- *Causal-ICM* provides principled uncertainty quantification across the full support of the target population. Crucially, we theoretically show that *Causal-ICM* limits the amount of information learnt from the observational dataset, and safeguards against overconfidence in the presence of bias.

- Through a comprehensive set of simulation studies and an application to a real-world dataset, we demonstrate that *Causal-ICM* achieves similar or superior performance both in point estimation and uncertainty quantification relative to a broad set of state-of-the-art causal data fusion methods.

2. Related Literature

Heterogeneous treatment effect (HTE) estimation has received renewed interest in recent years. The most common strategy is to focus on the conditional average treatment effect (CATE) function, which quantifies the expected treatment effect given particular covariate values. Under suitable identifiability assumptions (Dahabreh and Hernán, 2019), the CATE can be estimated parametrically - e.g. using linear regression - or non-parametrically, e.g. using nearest-neighbour matching, kernel methods (Brantner et al., 2023), or tree-based methods including causal random forests (Wager and Athey, 2018) and their Bayesian counterpart (Hahn et al., 2020). A related strand of work studies semi-parametric and debiased machine learning approaches that combine flexible outcome models with orthogonal score constructions to obtain robust estimates under high-dimensional confounding.

Data fusion across an RCT and an observational study has tended to focus on ATE estimation (see Lin et al. (2024) and Colnet et al. (2024) for comprehensive reviews). These methods aim to improve efficiency by integrating information from both sources. Demirel et al. (2024) focus instead on data fusion for generalisability, proposing a framework that builds an observational predictor and then treat the difference between the trial outcome function and that observational predictor as a bias function that captures both confounding and transportation bias due to the differences between the trial and target populations. Similar approaches, although outside the immediate scope of our work, have focused on data fusion for the estimation of long term effects, where randomised data are unconfounded but contain only short term effects, while observational data, although confounded, contain information about long term outcomes (Ghassami et al. (2025), Imbens et al. (2025)).

Data fusion methods for CATE estimation, the focus of this paper, follow three broad lines of work. One strategy is to use experimental data to debias observational estimates by learning a correction or bias function. Kallus et al. (2018) introduce a two-step approach that corrects hidden confounding in the absence of covariate overlap, relying on the strong assumption of linear confounding and the ability to identify this correction term parametrically. Yang et al. (2025b) relax this linearity assumption and provide identifiability and efficiency results under more flexible structural models. Hatt et al. (2022) develop a representation-learning strategy that learns shared features and confounded outcome models from observational data, using the RCT to estimate a bias function that debiases these models, with finite-sample bounds that highlight how performance depends on sample size, distribution shift, and bias complexity. Wu and Yang (2022) propose an R-learner that achieves consistency and asymptotic efficiency under covariate overlap between the RCT and the observational study. A second direction consists of "test-then-pool" strategies, in which the trial and observational datasets are combined only if the observational study appears sufficiently unbiased. For instance, Yang et al. (2023) implement such a strategy, although their approach is limited to linear treatment effect models and relies on strong parametric assumptions. A third line of work uses Bayesian dynamic borrowing to regulate how much information is extracted from the observational study. For instance, Lin et al. (2025) propose a Bayesian dynamic borrowing approach through a power likelihood to reduce the effect of confounding and control the degree of information we borrow from the observational study.

Our method, *Causal-ICM*, takes a Bayesian dynamic borrowing approach based on multi-task GPs, whereby each potential outcome is treated as a distinct task. Multi-task GPs

have been successfully employed for causal inference using solely observational data (Alaa and van der Schaar, 2017) and extended to handle hierarchical hidden confounders (Witty et al., 2020). Others have built on this framework to develop a general class of counterfactual multi-task deep kernel models that efficiently estimate causal effects and learn policies by stacking coregionalized GPs and deep kernels (Caron et al., 2022). GPs have also been employed as a matching tool for causal inference (Huang et al., 2023). Recent work has also explored Gaussian-process-based partially linear models for heterogeneous treatment effects, combining parametric effect components with flexible GP nuisance estimation (Horii and Chikahara, 2024). However, these approaches focus on single-source observational settings. In contrast, our work develops a multi-task GP framework specifically designed for causal data fusion between experimental and observational sources. Our approach is most similar to concurrent work employing multi-task GPs for pseudo-outcome regression to obtain error bounds on the observational bias term (Fawkes et al., 2025).

3. Methodology

We begin with a brief introduction on data fusion for HTE estimation. We then introduce a multi-task GPs framework tailored for this purpose. Throughout the paper, we use bold-face to denote vectors, e.g. $\mathbf{x}$, and capitals to denote matrices, e.g. X.

3.1. Data Fusion for Heterogeneous Treatment Effect Estimation

In this work, we focus on combining information from an observational dataset and a RCT (experimental) dataset. We begin with some notation. Let $S \in \{\mathrm{o}, \mathrm{e}\}$ indicate the study, with 'e' denoting the experimental sample and 'o' the observational one. Let $A \in \{0, 1\}$ be the binary treatment indicator. We observe the same baseline covariates in both studies, so each individual has $\mathbf{X} \in \mathcal{X}^S \subseteq \mathbb{R}^p$, where $\mathcal{X}^{\mathrm{e}}$ and $\mathcal{X}^{\mathrm{o}}$ overlap but $\mathcal{X}^{\mathrm{o}}$ is not necessarily contained in $\mathcal{X}^{\mathrm{e}}$. Our methodology requires overlap between the covariate distributions of the RCT and the observational study, but it does not require the support of the RCT covariates to be fully contained within that of the observational study. In regions with little or no randomized evidence, the observational study may still provide useful information for generalising treatment-effect estimates, although uncertainty should increase accordingly.

Without loss of generality, we distinguish three possible roles for baseline covariates: treatment-effect modifiers $\mathbf{X}_\tau \in \mathcal{X}^{\mathrm{e}}_\tau$, confounders $\mathbf{X}_W \in \mathcal{X}^{\mathrm{e}}_W$, and selection variables $\mathbf{X}_S \in \mathcal{X}^{\mathrm{e}}_S$, such that $\mathbf{X} = \mathbf{X}_\tau \cup \mathbf{X}_W \cup \mathbf{X}_S$. Here, confounders are variables that act as common causes of treatment and outcome, treatment-effect modifiers are variables indexing the heterogeneity of interest, and selection variables are variables whose distribution may differ between the RCT and the target population. These roles are not assumed to be mutually exclusive. Finally, we observe a continuous outcome $Y \in \mathbb{R}$. The counterfactual outcome Y_i^a denotes the value individual i would have had under treatment $A = a$. The causal estimand of interest is the CATE in the target population given the baseline covariates X, defined for $\mathbf{x}_\tau \in \mathcal{X}^{\mathrm{e}} \cup \mathcal{X}^{\mathrm{o}}$ as

$$\tau(\mathbf{x}_\tau) = \mathbb{E}\left[Y^{a=1} - Y^{a=0} | \mathbf{X}_\tau = \mathbf{x}_\tau\right]$$

The difference in the conditional expectations of the outcomes in each study is denoted

$$\omega^s(\mathbf{x}_\tau) = \mathbb{E}\left[Y | \mathbf{X}_\tau = \mathbf{x}_\tau, A = 1, S = s\right] - \mathbb{E}\left[Y | \mathbf{X}_\tau = \mathbf{x}_\tau, A = 0, S = s\right].$$

The fundamental problem of causal inference is that for each unit we cannot observe both potential outcomes simultaneously. Hence, we make the following identifiability assumptions, which are standard in the literature (e.g. Lanners et al. (2025), Shi et al. (2023)):

A1 Consistency: $A = a \Rightarrow Y = Y^a$

A2 Mean conditional exchangeability over treatment for the treatment contrast given $\mathbf{X}_\tau$ in the RCT:

$$\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau, S = \mathrm{e}] = \mathbb{E}[Y^1 \mid A = 1, \mathbf{X}_\tau, S = \mathrm{e}] - \mathbb{E}[Y^0 \mid A = 0, \mathbf{X}_\tau, S = \mathrm{e}]$$

A3 Positivity of treatment assignment in the RCT: $P(A = a|\mathbf{X}_S = \mathbf{x}_S, S = \mathrm{e}) > 0$ for $\mathbf{x}_S \in \mathcal{X}_S^{\mathrm{e}}$.

A4 Mean conditional exchangeability over selection into the RCT for the treatment contrast given $\mathbf{X}_\tau$, $\mathbf{X}_S$: $\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau \cup \mathbf{X}_S] = \mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}]$

A5 Positivity of RCT participation: $P(S = \mathrm{e}|\mathbf{X}_S = \mathbf{x}_S) > 0$ for $\mathbf{x}_S \in \mathcal{X}_S^{\mathrm{e}}$

A6 Conditional invariance of selection variable distributions between the RCT and the target population: $P(X_S|X_\tau) = P(X_S|X_\tau, S = \mathrm{e})$.

Under assumptions A1 - A6 the CATE can be identified from the RCT data as $\tau(\mathbf{x}_\tau) = \omega^{\mathrm{e}}(\mathbf{x}_\tau)$ for $\mathbf{x}_\tau \in \mathcal{X}_\tau^{\mathrm{e}}$. The derivation is provided in Appendix A, and a discussion on Assumptions A4 and A6 is provided on Appendix B. For notational simplicity, and with a slight abuse of notation, we henceforth use $\mathbf{X}$ and $\mathbf{x}$ to denote whichever subset of covariates is relevant in context, including treatment-effect modifiers, confounders, or selection covariates.

Conversely, relative to the RCT, the covariate distribution of the observational study may be more representative of the target population. In the presence of unmeasured confounding, however, it is not possible to identify $\tau(\mathbf{x})$ from the observational data alone. In particular, the hidden confounding effect in the observational data, denoted by $\eta(\mathbf{x}) = \tau(\mathbf{x}) - \omega^{\mathrm{o}}(\mathbf{x})$, is nonzero. As a result, estimates of $\omega^{\mathrm{o}}(\mathbf{x})$ may be biased for the estimand $\tau(\mathbf{x})$.

In this work, we use a multi-task Gaussian process to jointly estimate $(\omega^{\mathrm{o}}(\mathbf{x}), \omega^{\mathrm{e}}(\mathbf{x}))$, treating the two data sources as separate tasks. This leverages both the unbiasedness of the RCT and the broader covariate support of the observational study, while adaptively accounting for confounding by comparing $(\omega^{\mathrm{o}}(\mathbf{x}), \omega^{\mathrm{e}}(\mathbf{x}))$ within $\mathcal{X}^{\mathrm{e}}$. Unlike other methods, we avoid strong parametric assumptions about the bias function, relying instead on the multi-task GP to extrapolate non-linearly with posterior uncertainty quantification, so that regions with low covariate support naturally exhibit higher uncertainty.

3.2. Multi-task Gaussian Processes

A GP is a collection of random variables where any finite subset have a joint Gaussian distribution. In the scalar case, the distribution of a GP f is completely specified by its mean function $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. We will assume that $m(\mathbf{x}) = 0$ throughout. Given a set of observations $(y_i, \mathbf{x}_i)_{i=1}^n$, we seek to learn a function f such that $y_i = f(\mathbf{x}_i) + \epsilon_i$, where we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. GP regression proceeds by placing a GP prior on the function f, and computing the posterior distribution of f given the observations.

In our setting, we use a T-learner approach (Künzel et al., 2019) to estimate $\tau(\mathbf{x})$, i.e. we individually estimate $\mathbb{E}[Y^0 \mid \mathbf{x}]$ and $\mathbb{E}[Y^1 \mid \mathbf{x}]$ with separate models. For brevity, we will focus on $\mathbb{E}[Y^1 \mid \mathbf{x}]$ in the exposition. We write the study specific response surfaces as:

$$f^{\mathrm{e}}(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{x}, S = \mathrm{e}, A = 1], \quad f^{\mathrm{o}}(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{x}, S = \mathrm{o}, A = 1].$$

For $\mathbf{f}(\mathbf{x}) = (f^{\mathrm{e}}(\mathbf{x}), f^{\mathrm{o}}(\mathbf{x}))$, we assign a multi-task GP prior $\mathbf{f} \sim \mathcal{GP}(\boldsymbol{\mu}, K)$, where $\boldsymbol{\mu}$ is a bivariate mean function (taken again to be 0) and K is a positive 2×2 *matrix-valued* covariance function that maps input points $\mathbf{x}_i$ and $\mathbf{x}_j$ to matrices quantifying the relationship of the outputs at these inputs. We specify K in the next section. The key strength of multi-task GPs is the ability to share information between tasks, which we will leverage for data fusion. See Appendix C Fig 3 for an illustrative comparison to independent GPs.

We assume a Gaussian observation model,

$$y_i^{\mathrm{e}} = f^{\mathrm{e}}(\mathbf{x}_i^{\mathrm{e}}) + \epsilon_i^{\mathrm{e}}, \quad y_j^{\mathrm{o}} = f^{\mathrm{o}}(\mathbf{x}_j^{\mathrm{o}}) + \epsilon_j^{\mathrm{o}},$$

where $\epsilon_i^{\mathrm{e}}, \epsilon_j^{\mathrm{o}} \sim \mathcal{N}(0, \sigma^2)$ independently across tasks and observations for $i = 1, \ldots, n^{\mathrm{e}}$ and $j = 1, \ldots, n^{\mathrm{o}}$. For simplicity, we assume a common variance σ^2 between tasks but this can be easily extended to task-specific variances. We write $\mathbf{y}^{\mathrm{e}}$ for the n^{e}-vector of outcomes and X^{e} for the $(n^{\mathrm{e}} \times p)$ matrix of covariates respectively, and similarly for the observational dataset.

In data fusion settings for clinical trials, typically $n^{\mathrm{e}} \ll n^{\mathrm{o}}$, and $p(\mathbf{x}^{\mathrm{e}})$ may not have support over the entire target population. Making inferences across the whole population thus requires some level of extrapolation. We leverage the multi-task GP to improve estimation of f^{e} outside the support of $p(\mathbf{x}^{\mathrm{e}})$, given the observations $(\mathbf{y}^{\mathrm{o}}, X^{\mathrm{o}})$. By sharing information between tasks and leveraging the flexible nature of GPs, we hope to increase precision of our estimates without introducing significant bias. The object of interest is the posterior distribution of f^{e} given $\mathcal{D} = (\mathbf{y}^{\mathrm{e}}, X^{\mathrm{e}}, \mathbf{y}^{\mathrm{o}}, X^{\mathrm{o}})$, which, by conjugacy, is also a GP. Crucially, the Bayesian framework also provides reliable uncertainty estimates in regions of extrapolation.

3.3. Intrinsic coregionalisation models (ICMs)

The Intrinsic Coregionalization Model (ICM) is a special case of a multi-task GP with a separable kernel, which is particularly interpretable and suitable for data fusion (Vargas-Guzman and Warrick, 1999). Each task is a linear combination of independent latent functions, yielding a particularly simple covariance structure. For our setting, we have

$$f^{\mathrm{e}}(\mathbf{x}) = \sum_{q=1}^{Q} \alpha_q^{\mathrm{e}} u_q(\mathbf{x}), \quad f^{\mathrm{o}}(\mathbf{x}) = \sum_{q=1}^{Q} \alpha_q^{\mathrm{o}} u_q(\mathbf{x}),$$

where $u_q(\mathbf{x})$ are independent zero-mean latent functions taken from the same scalar GP, $u_q \sim \mathcal{GP}(0, k)$ for a chosen kernel k. The scalar coefficients $(\alpha_q^{\mathrm{e}}, \alpha_q^{\mathrm{o}})$ are which will be used to construct the coregionalization matrix defined shortly, which governs the dependence structure. The rank, Q, of the ICM is the number of independent components, which we set to $Q = 2$ for the single arm case; we provide a discussion on the interpretation of this later.

The multi-task covariance function can be expressed as

$$K(\mathbf{x}, \mathbf{x}') = B \otimes k(\mathbf{x}, \mathbf{x}'),$$

where $\otimes$ is the Kronecker product and B is the coregionalization matrix taking values

$$B = \begin{bmatrix} \beta^{\mathrm{e}} & \beta^{\mathrm{eo}} \\ \beta^{\mathrm{eo}} & \beta^{\mathrm{o}} \end{bmatrix} = \begin{bmatrix} (\alpha_1^{\mathrm{e}})^2 + (\alpha_2^{\mathrm{e}})^2 & \alpha_1^{\mathrm{e}}\alpha_1^{\mathrm{o}} + \alpha_2^{\mathrm{e}}\alpha_2^{\mathrm{o}} \\ \alpha_1^{\mathrm{e}}\alpha_1^{\mathrm{o}} + \alpha_2^{\mathrm{e}}\alpha_2^{\mathrm{o}} & (\alpha_1^{\mathrm{o}})^2 + (\alpha_2^{\mathrm{o}})^2 \end{bmatrix}. \tag{1}$$

The posterior distribution over f^{e} can then be computed using standard theory (see Appendix D). For a given test point $\mathbf{x}_*$, we have that the posterior distribution is Gaussian:

$$[f^{\mathrm{e}}(\mathbf{x}_*) \mid \mathcal{D}] \sim \mathcal{N}(m^{\mathrm{e}}(\mathbf{x}_*), V^{\mathrm{e}}(\mathbf{x}_*)),$$
$$m^{\mathrm{e}}(\mathbf{x}_*) = k^{\mathrm{e}}(\mathbf{x}_*, X)\,\Sigma^{-1}\mathbf{y}, \quad V^{\mathrm{e}}(\mathbf{x}_*) = \beta^{\mathrm{e}} k(\mathbf{x}_*, \mathbf{x}_*) - k^{\mathrm{e}}(\mathbf{x}_*, X)\,\Sigma^{-1} k^{\mathrm{e}}(X, \mathbf{x}_*).$$

Here, $\mathbf{y} = (\mathbf{y}^{\mathrm{e}}, \mathbf{y}^{\mathrm{o}})$ and $X = (X^{\mathrm{e}}, X^{\mathrm{o}})$ are the concatenated response vectors and covariate matrices respectively, and $\Sigma = K(X, X) + \sigma^2 I$ with

$$K(X, X) = \begin{bmatrix} \beta^{\mathrm{e}} K(X^{\mathrm{e}}, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} K(X^{\mathrm{e}}, X^{\mathrm{o}}) \\ \beta^{\mathrm{eo}} K(X^{\mathrm{o}}, X^{\mathrm{e}}) & \beta^{\mathrm{o}} K(X^{\mathrm{o}}, X^{\mathrm{o}}) \end{bmatrix},$$

where $K(X^s, X^t)$ is the regular (cross-)covariance matrix of the kernel $k(\mathbf{x}, \mathbf{x}')$ for the covariate matrices from $S = s$ and $S = t$. We similarly have

$$k^{\mathrm{e}}(\mathbf{x}_*, X) = \begin{bmatrix} \beta^{\mathrm{e}} k(\mathbf{x}_*, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{o}}) \end{bmatrix},$$

where $k(\mathbf{x}_*, X^s)$ is the row vector of covariances between $\mathbf{x}_*$ and X^s, and $k^{\mathrm{e}}(X, \mathbf{x}_*) = k^{\mathrm{e}}(\mathbf{x}_*, X)^T$. Given this posterior, $m^{\mathrm{e}}(\mathbf{x}_*)$ provides point estimates of the response surface, with associated posterior variance $V^{\mathrm{e}}(\mathbf{x}_*)$ which can be used to compute credible intervals.

3.4. *Causal-ICM*

We now introduce *Causal-ICM*. Specifically, we propose an interpretable parametrisation of the ICM for causal inference and introduce a bespoke procedure for learning its coefficients, rather than relying on standard marginal likelihood maximisation. This serves to control the influence of the observational dataset and avoid a situation in which a large confounded observational sample yields biased estimates with artificially narrow uncertainty intervals relative to the RCT. *Causal-ICM* consists of the following choices for the ICM coefficients:

$$\begin{bmatrix} \alpha_1^{\mathrm{e}} & \alpha_2^{\mathrm{e}} \\ \alpha_1^{\mathrm{o}} & \alpha_2^{\mathrm{o}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$$

Under this parameterization, $\rho \in (0, 1)$ controls the degree of borrowing, with $\rho \to 1$ corresponding to maximum borrowing and $\rho \to 0$ to none (We provide an additional interpretation of ρ in Appendix E, Proposition 2). Although this implies $\beta^{\mathrm{e}} = \beta^{\mathrm{o}} = 1$ (and $\beta_{\mathrm{eo}} = \rho$), the model remains flexible as long as the kernel $k(\mathbf{x}, \mathbf{x})$ includes an independent scaling term (e.g., the variance in an RBF kernel) and the scales of f^{e} and f^{o} are similar.

With these suggested settings for the ICM coefficients, we then have:

$$f^{\mathrm{e}}(\mathbf{x}) = u_1(\mathbf{x}), \quad f^{\mathrm{o}}(\mathbf{x}) = \rho f^{\mathrm{e}}(\mathbf{x}) + \sqrt{1-\rho^2}\, u_2(\mathbf{x}).$$

Thus, $f^{\mathrm{e}} = u_1$ and f^{o} is a scaled version of f^{e} plus a term capturing confounding. Setting $\alpha_2^{\mathrm{e}} = 0$ means that the experimental regression surface is modelled directly as a single

GP, while the observational regression surface decomposes into a shared component and an observational-specific component. Choosing $\rho \in (0,1)$ is reasonable, as $f^{\mathrm{o}}(\mathbf{x})$ is expected to be positively correlated with $f^{\mathrm{e}}(\mathbf{x})$, both model the conditional expectation of the outcome, with f^{o} additionally contaminated by confounding. In practice, restricting α_2^{e} minimally affects flexibility while improving interpretability. Remaining hyperparameters, including kernel parameters and sampling variance, are estimated via marginal likelihood maximisation.

3.4.1. Variance bound for conditional mean function under *Causal-ICM*

When evaluating treatment effects, it is important to quantify uncertainty alongside point estimates. The multi-task Gaussian process provides this directly via the posterior variance $V^{\mathrm{e}}(\mathbf{x}_*)$. In data fusion, a key concern is that large observational samples may lead to biased estimates with unrealistically narrow uncertainty intervals when unmeasured confounding is present. *Causal-ICM* mitigates this, as shown by a lower bound on the posterior variance $V^{\mathrm{e}}(\mathbf{x}_*)$ even as n^{o} grows. This variance result can be viewed as formally limiting the information flow from the observational dataset, similar to cut models (Lin et al., 2025).

Proposition 1 *Suppose $\mathbf{x}_*$ is a test point of interest. Let $V^{\mathrm{e}}(\mathbf{x}_*)$ and $V^{\mathrm{e}}_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*)$ denote the posterior variances of $f^{\mathrm{e}}(\mathbf{x}_*)$ given the full dataset $\mathcal{D} = (\mathbf{y}^{\mathrm{e}}, X^{\mathrm{e}}, \mathbf{y}^{\mathrm{o}}, X^{\mathrm{o}})$ and the experimental dataset $\mathcal{D}_{\mathrm{e}} = (\mathbf{y}^{\mathrm{e}}, X^{\mathrm{e}})$ only respectively. The posterior variance given $\mathcal{D}$ then satisfies*

$$V^{\mathrm{e}}\left(\mathbf{x}_*\right) \geq \left(1-\rho^2\right) V^{\mathrm{e}}_{\mathcal{D}_{\mathrm{e}}}\left(\mathbf{x}_*\right) \tag{2}$$

where $\rho \in (0,1)$ is the borrowing hyperparameter.

Proof Let $n^{\mathrm{e}}, n^{\mathrm{o}} \geq 0$ denote the sizes of the experimental and observational datasets respectively. We outline the proof for $n^{\mathrm{e}} = 0$ here, which corresponds to the prior-only case for the RCT. The proof for the general case $n^{\mathrm{e}} > 0$ is deferred to Appendix F. The key is that when $n^{\mathrm{e}} = 0$, the posterior variance can be written as

$$V^{\mathrm{e}}\left(\mathbf{x}_*\right) = \left(1-\rho^2\right) k(\mathbf{x}_*, \mathbf{x}_*) + \rho^2 V^{\mathrm{e}}_{\mathcal{D}_{\mathrm{o}}}\left(\mathbf{x}_*\right)$$

where $V^{\mathrm{e}}_{\mathcal{D}_{\mathrm{o}}}$ is the posterior variance for a GP with kernel $k(\mathbf{x}, \mathbf{x})$ fit only to $\mathcal{D}_{\mathrm{o}} = (\mathbf{y}^{\mathrm{o}}, X^{\mathrm{o}})$, and $k(\mathbf{x}_*, \mathbf{x}_*)$ is the prior variance. As $V^{\mathrm{e}}_{\mathcal{D}_{\mathrm{o}}}\left(\mathbf{x}_*\right) \geq 0$, we obtain the desired lower bound. ■

Proposition 1 shows that the Causal-ICM posterior variance is lower bounded by a non-degenerate fraction of the experimental-only posterior variance. The same decomposition also yields an upper bound. Writing

$$V^{\mathrm{e}}_D(x^*) = V^{\mathrm{e}}_{D_{\mathrm{e}}}(x^*) - \rho^2\big(k(x^*, X^{\mathrm{o}}) - k'(x^*, X^{\mathrm{o}})\big)\Sigma^{-1}_{eo}\big(k(X^{\mathrm{o}}, x^*) - k'(X^{\mathrm{o}}, x^*)\big),$$

and noting that the quadratic form is non-negative, we obtain

$$(1-\rho^2)V^{\mathrm{e}}_{D_{\mathrm{e}}}(x^*) \leq V^{\mathrm{e}}_D(x^*) \leq V^{\mathrm{e}}_{D_{\mathrm{e}}}(x^*).$$

Hence, incorporating observational data cannot inflate the posterior variance beyond the RCT-only baseline, nor arbitrarily concentrate the posterior when $\rho \in (0,1)$.

An intuitive understanding of the result is as follows: for two random variables with correlation ρ, knowledge of one random variable does not inform us of the exact value of

the other unless $\rho = 1$. The key interpretation of Proposition 1 in the data fusion context is as follows: even as $n^{\mathrm{o}} \to \infty$, the posterior variance of $f^{\mathrm{e}}(\mathbf{x}_*)$ conditional on the full dataset will at most decrease by a factor of $(1 - \rho^2)$ relative to the posterior variance given the experimental data only, where $|\rho| \leq 1$. We conjecture that this inequality is tight as $n^{\mathrm{o}} \to \infty$. This can be interpreted as limiting the 'effective sample size' of the observational dataset to be $1/\left(1 - \rho^2\right)$ relative to the experimental dataset. Crucially, this variance bound protects from having very tight credible intervals centred around a biased estimate - the posterior uncertainty will accurately reflect our distrust of the observational dataset. We also expect this 'effective sample size' effect to hold for the posterior mean of $f^{\mathrm{e}}(\mathbf{x}_*)$, where there the bias introduced to the posterior mean by $\mathcal{D}_{\mathrm{o}}$ is limited again by the choice of ρ. We leave this interesting direction for future work. Although our focus is on the conditional mean functions, our approach also provides the posterior distribution of the confounding function, offering additional insights that may be of independent interest (see Appendix G).

3.4.2. Tuning ρ

We now outline a data-adaptive procedure to select ρ, where the goal is to prevent information from the large confounded observational study from swamping that of the smaller unconfounded RCT. Our proposal is based on cross-validation to minimize the Root Mean Squared Error (RMSE) on weighted held-out RCT patients only, where the weighting tailors ρ for extrapolation. Specifically, we use 5-fold cross-validation, as the RCT sample size is usually small, and we consider $\rho \in [0.0, 0.1, \cdots, 1.0]$ on a grid. Our objective is:

$$\mathcal{L}(\rho) = \sum_{i=1}^{n_{\text{held-out}}} w(\tilde{\mathbf{x}}_i) \left(\tilde{y}_i - m^{\mathrm{e}}(\tilde{\mathbf{x}}_i)\right)^2, \quad w(\mathbf{x}) = \frac{1}{1 - p(S = \mathrm{o} \mid \mathbf{x})}$$

where $\{\tilde{y}_i, \tilde{\mathbf{x}}_i\}_{i=1}^{n_{\text{held-out}}}$ is a held-out subset of $(\mathbf{y}^{\mathrm{e}}, X^{\mathrm{e}})$ that was not used to train m^{e}. The above objective is motivated as follows. Firstly, we only evaluate predictions on RCT patients where there is no confounding present. Secondly, the weights will upweight RCT patients who have higher probability of arising from the observational dataset, so ρ will be tailored to improve extrapolation beyond the RCT support. We then expect the largest gains in power of ρ is chosen close to 1, indicating that the confounding effect is small. We will see in the experiments that this works well in practice.

3.4.3. Extension to CATE Estimation

When both treatment and control arms are available, we fit two separate rank-2 ICMs to model $(f_0^{\mathrm{e}}, f_0^{\mathrm{o}})$ and $(f_1^{\mathrm{e}}, f_1^{\mathrm{o}})$, focusing on sharing information between the experimental and observational datasets. This yields independent posteriors for f_0^{e} and f_1^{e}, which we combine to obtain the posterior of $\tau(\mathbf{x})$ by differencing the means and summing the variances. We allow distinct kernel hyperparameters for the two ICMs while using a common ρ, chosen to minimise the average $\mathcal{L}(\rho)$ across treatment groups. This setup corresponds to a T-learner approach (Künzel et al., 2019), and avoids imposing strong assumptions on the bias structure across treatment arms and preserves theoretical tractability. An alternative approach would be to also couple across treatment arms; however, in preliminary experiments not presented here, we found that such strongly coupled models resulted in over-shrinkage of posterior variance and reduced coverage.

4. Experiments

We investigated *Causal-ICM*'s performance through multiple simulation studies and a comprehensive analysis of Real-World Data (RWD) sourced from the Tennessee STAR study (Achilles et al., 2008). The code is available online[1]. The optimal value of ρ is chosen data-adaptively in all cases (Section 3.4.2). Model evaluation and comparison were based on the RMSE, averaged over the covariate distribution of the observational study.

We compared *Causal-ICM* against several benchmarks: (i) a T-learner with GPs regressors trained separately on each study, (ii) the two-step method of Kallus et al. (2018), which debiases observational data via a low-complexity bias function with GPs as base learners, (iii) the integrative estimator of Yang et al. (2025b), which assumes a non-linear CATE and linear confounding effect, (iv) the power likelihood method of Lin et al. (2025), and (v) the test-then-pool approach of Yang et al. (2023). For Yang et al. (2025b), we used the default settings recommended in their paper, while for Lin et al. (2025), we selected the power parameter η by maximizing the Expected Log Pointwise Predictive Density (ELPD), which in both cases resulted in $\eta = 0$ (no pooling). Finally, we assessed the coverage of *Causal-ICM*'s CATE credible intervals across simulation scenarios, comparing them with those from the T-learner approaches.

We used GPy (2012) to train the ICMs, with ρ tuned as described in Section 3.4.2 and remaining hyperparameters estimated via marginal likelihood maximisation. All experiments were conducted on a MacBook Pro (2022) equipped with an Apple M2 chip, 8 CPU cores, 10 GPU cores, and 16 GB of memory, running macOS 26.2. Runtime comparisons indicate that *Causal-ICM* incurs a moderate computational overhead relative to single-study GP baselines, but remains competitive with other data-fusion approaches (see Appendix K).

4.1. Simulation Studies

We showcase the performance of *Causal-ICM* in the main text using two distinct univariate simulation studies. Multivariate simulation studies, are detailed in Appendix H, illustrating the competitive performance of *Causal-ICM* in higher dimensions. In all simulations, the sample size is roughly between 200-300 for the RCT and 1000 for the observational study.

In the first simulation setting, we assume a continuous baseline covariate $X \sim \text{Unif}\,[-2, 2]$. The values of X control the trial participation probability, where $S \sim \text{Bernoulli}\,(p_S)$ for $p_S = \exp\,(-3 - 3X)/(1 + \exp\,(-3 - 3X))$. Here, patients with smaller X values have higher probability of being assigned in the trial. Trial participants are randomly assigned treatment with $A \sim \text{Bernoulli}\,(0.5)$. The observed outcomes are generated as $Y = A\tau\,(X) + X + \epsilon$ for $A \in \{0, 1\}$, where $\tau\,(X) = 1 + X$ is the CATE and $\epsilon \sim \mathcal{N}\,(0, 1)$. For the observational study we have $X \sim \text{Unif}\,[-2, 2]$, with the treatment generated according to $A \sim \text{Bernoulli}\,(e\,(X))$ where $\text{logit}\,(e\,(X)) = -X$. The observed outcomes are $Y = A\tau\,(X) + X + U + \epsilon$ for $A \in \{0, 1\}$. Here, the hidden confounder is generated from $U \sim \mathcal{N}\,((2A - 1)\,X, 1)$, yielding a linear confounding effect $\eta(X) = 2X$. In the second simulation, we have non-linear potential outcomes, CATE and confounding effect, as follows: $Y = A\tau(X) + X^2 - 1 + U + \epsilon$, where $\tau(X) = 1 + X + X^2$ and $U \sim \mathcal{N}((2A - 1)\sin(X - 1), 1)$, yielding $\eta(X) = 2\sin(X - 1)$. All other variables remain the same.

1. https://github.com/EvanDimitriou/CausalICM

Under a linear CATE and a linear unobserved confounding effect, *Causal-ICM* with $\rho = 0.1$ exhibited comparable performance to the integrative HTE (Yang et al. (2025b)) and Elastic HTE estimators (Yang et al. (2023)), and outperformed all other techniques (Fig 1 left). In the non-linear setting (both CATE and unobserved confounding effect), *Causal-ICM* outperformed all other methods yielding the lowest RMSE (Fig 1 right). These results reflect the limitations of the experimental grounding, the Integrative HTE and the Elastic HTE methods. Each of these are designed for linear treatment or confounding effect, resulting in inflated RMSE values in scenarios where these assumptions do not hold.

Regarding uncertainty quantification, our approach generated 95% credible intervals with reasonable, though slightly conservative, coverage close to the nominal level in both simulation settings (Fig. 2). This conservativeness is consistent with the structure of the model: posterior uncertainty reflects not only sampling variability, but also uncertainty induced by partial cross-domain borrowing and the possibility of residual observational bias. The observational-only T-learner had the lowest coverage across the baseline covariate support, which is expected in the presence of unmeasured confounding: as sample size grows, its estimates can remain biased while posterior uncertainty shrinks, leading to overconfident credible intervals that fail to cover the true CATE. This is exactly the behaviour that Causal-ICM is robust to, as formalized in Proposition 1: the large sample size of the observational study does not overwhelm the *Causal-ICM* estimates. See Figure 5 in Appendix I for an illustrative example of extrapolation and uncertainty quantification. Additional sensitivity analysis regarding imbalance of study sample sizes, performance within and outside the RCT support, kernel choice and the degree of overlap between the two studies and the choice of ρ is provided in Appendix J.

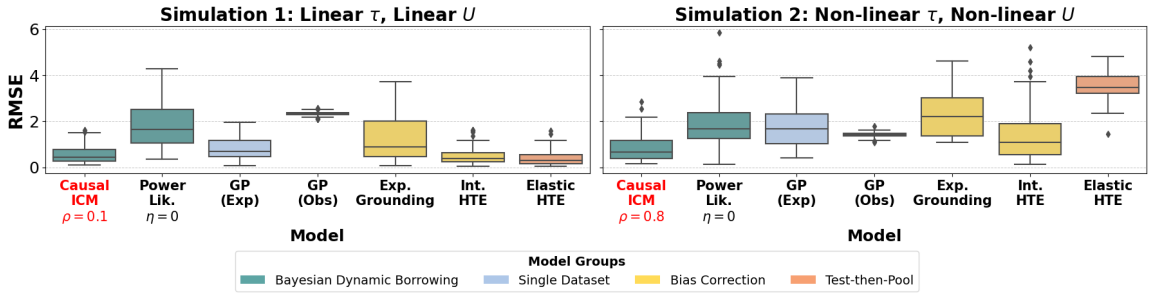

Figure 1: Simulation results over 100 simulated datasets. *Left*: Simulation setting 1: linear CATE and unobserved confounding, *Right*: Non-linear CATE and unobserved confounding

4.2. Real-World Data Analysis

The Tennessee Student/Teacher Achievement Ratio (STAR) (Achilles et al., 2008) randomised experiment initiated in 1985 studied the impact of class size on student outcomes through standardized test scores from first to third grade. We focus on two 'treatments', here corresponding to two experimental conditions: small class size and regular class size. We followed a similar approach to Kallus et al. (2018) and used the real data to produce a 'confounded' dataset, corresponding to the observational sample, and a smaller unconfounded dataset, which comprised our RCT data. After removing subjects with missing treatment or outcome values, we were left with a randomised sample of 4218 students. Baseline covariates

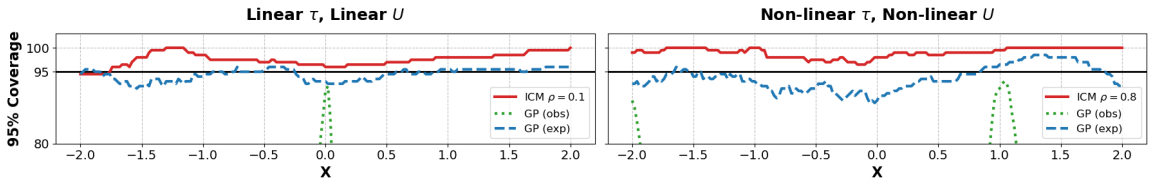

Figure 2: 95% conditional coverage over the covariate distribution of the observational study for the two simulation studies. The black solid line indicates the nominal level (95%). *Left*: First simulation setting (linear CATE and confounding), *Right*: Second simulation setting (non-linear CATE and confounding). The GP (obs) method exhibited 0% coverage in most of the support covariate distribution.

included gender, race, birth month, birth year, free lunch eligibility and teacher ID. To obtain the 'experimental' dataset we sampled randomly only from rural or inner city students. To form the confounded observational dataset, we aggregated all control cases not included in the experimental data. Additionally, for those undergoing treatment, we drew a sample with a down-weighting mechanism applied to individuals whose outcomes fell below the 30^{th} percentile. After preprocessing, the unconfounded dataset included 422 students, while the confounded dataset included 2593. Further, 379 students were kept as a validation set. We compared *Causal-ICM* (optimal $\rho = 0.4$) with a GP based T-learner trained either on the experimental or the observational data,the experimental grounding method of Kallus et al. (2018) either with Random Forest or GP as base learners and the Integrative HTE method of Yang et al. (2025b). We were unable to include the Elastic HTE estimator (Yang et al. (2023)) or the power likelihood approach (Lin et al. (2025)), as both methods encountered convergence issues due to the datasets' high dimensionality and the limited variability in their categorical covariates.

In order to compare *Causal-ICM* with existing methods, we require a notion of ground truth. Given the randomised nature of the study, we assume that an unbiased estimate of the CATE can be obtained using a doubly robust estimator, which we then treat as the ground truth. To obtain this, we estimated the propensity score and conditional expectation models using the full dataset, prior to splitting it into the confounded and unconfounded samples (Saito and Yasui, 2020). We note that, in the absence of the true CATE, other performance metrics adapted to the counterfactual nature of the evaluation are also available (Alaa and Van Der Schaar, 2019; Boyer et al., 2023). The results are summarised in Table 1. *Causal-ICM* again performed highly competitively in terms of RMSE, on a par with the experimental grounding method with Random Forests as base learners.

5. Discussion

We have developed *Causal-ICM*, a rank-2 ICM to combine observational and experimental data to estimate treatment effects for a target population, where the hyperparameter ρ controls the degree of borrowing in an interpretable way. In contrast to existing methods, *Causal-ICM* is highly flexible and does not make assumptions about the functional form of the CATE or the effect of unobserved confounding. By construction, the method trades a small amount of potential bias from the observational source against gains in precision and

Table 1: RMSE values for the RWD example. *Causal-ICM*: our method; GP (exp): T-learner with GPs as base learners trained in the unconfounded; GP (obs): T-learner with GPs as base learners trained in the confounded data; Experimental grounding (GP): the method proposed by Kallus et al. (2018) with GPs as base learners; Experimental grounding (RF): the method proposed by Kallus et al. (2018) with Random Forest as base learners; Integrative HTE: the method proposed by Yang et al. (2025b).

Method	**RMSE**	**Method**	**RMSE**
Causal-ICM ($\rho = 0.4$)	**6.29**	Experimental grounding (GP)	6.38
GP (exp)	6.36	Experimental grounding (RF)	**6.27**
GP (obs)	7.47	Integrative HTE	11.84

external validity, while anchoring inference to the RCT. This trade-off is further supported by our additional experiments, which show that *Causal-ICM* remains stable under increasing observational sample sizes and continues to perform competitively even when alternative data-fusion approaches deteriorate. By providing accurate estimates of HTEs together with uncertainty quantification, *Causal-ICM* has the potential to inform more personalised medical decision making. By identifying regions of the population where treatment effects remain highly uncertain, our approach can help guide the design of future trials to improve representativeness and reduce uncertainty to satisfactory levels (Yang et al. (2025a)).

Despite the inherent difficulty of extrapolation, *Causal-ICM* performs well in these settings, as evidenced by our additional analysis separating performance within and outside the support of the randomized trial, where the gains are particularly pronounced in regions lacking experimental data. *Causal-ICM* also achieves reasonable, if slightly conservative, credible intervals. In general, it is challenging to ensure frequentist coverage for Bayesian nonparametric methods. The slightly conservative empirical coverage of *Causal-ICM* can be explained by the model hedging against disagreement between the experimental and observational signals, with additional uncertainty induced by partial cross-domain borrowing. A related limitation of our approach is the sensitivity of uncertainty estimates to the kernel choice and the hyperparameter optimization, itself an intrinsic challenge of GPs.

The flexibility of the (multi-task) GP framework underlying *Causal-ICM* opens several avenues for future work. In particular, the current model uses shared kernel hyperparameters across latent functions but more flexible alternatives, such as the Linear Model of Coregionalization (LCM; Alvarez et al. (2012)), could allow task-specific structure. Other possible extensions include incorporating more than two data sources or treatment arms, and placing a prior on the borrowing parameter ρ. More broadly, extending the framework to allow non-Gaussian likelihoods (e.g. for binary or heavy-tailed outcomes) is of significant practical interest, but will likely require non-trivial posterior approximations.

References

C.M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. Tennessee Student/Teacher Achievement Ratio (STAR) project, April 2008. URL https://doi.org/10.7910/DVN/SIWH9F.

Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 191–201. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/alaa19a.html.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6a508a60aa3bf9510ea6acb021c94b48-Abstract.html.

Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. doi: 10.1561/2200000036. URL https://doi.org/10.1561/2200000036.

Christopher B. Boyer, Issa J. Dahabreh, and Jon A. Steingrimsson. Assessing model performance for counterfactual predictions, 2023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.70287.

Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A. Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical Science*, 38(4):640–654, 2023. doi: 10.1214/23-STS890. URL https://doi.org/10.1214/23-STS890.

Alberto Caron, Ioanna Manolopoulou, and Gianluca Baio. Counterfactual learning with multioutput deep kernels. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=iGREAJdULX.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 2024. doi: 10.1214/23-STS889. URL https://doi.org/10.1214/23-STS889.

Issa J. Dahabreh and Miguel A. Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, August 2019. doi: 10.1007/s10654-019-00533-2. URL https://doi.org/10.1007/s10654-019-00533-2.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023. doi: 10.1146/annurev-statistics-042522-103837. URL https://doi.org/10.1146/annurev-statistics-042522-103837.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10385–10408. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/demirel24a.html.

Jake Fawkes, Michael O'Riordan, Athanasios Vlontzos, Oriol Corcoll, and Ciarán Mark Gilligan-Lee. The hardness of validating observational studies with experimental data. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1819–1827. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/fawkes25b.html.

Thomas R. Frieden. Evidence for health decision making — beyond randomized, controlled trials. *New England Journal of Medicine*, 377(5):465–475, 2017. doi: 10.1056/NEJMra1614394. URL https://doi.org/10.1056/NEJMra1614394.

AmirEmad Ghassami, Chang Liu, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. Combining experimental and observational data for identification and estimation of long-term causal effects, 2025. URL https://arxiv.org/abs/2201.10743.

GPy. GPy: A Gaussian process framework in Python, 2012. URL https://github.com/SheffieldML/GPy.

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, September 2020. ISSN 1936-0975, 1931-6690. doi: 10.1214/19-BA1195. Publisher: International Society for Bayesian Analysis.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects, 2022. URL https://arxiv.org/abs/2202.12891.

Shunsuke Horii and Yoichi Chikahara. Uncertainty quantification in heterogeneous treatment effect estimation with gaussian-process-based partially linear model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20420–20429, 2024. doi: 10.1609/aaai.v38i18.30025. URL https://doi.org/10.1609/aaai.v38i18.30025.

Bin Huang, Chen Chen, Jinzhong Liu, and Siva Sivaganisan. GPMatch: A Bayesian causal inference approach using Gaussian process covariance function as a matching tool. *Frontiers in Applied Mathematics and Statistics*, 9, 2023. ISSN 2297-4687. doi: 10.3389/fams.2023.1122114. URL https://www.frontiersin.org/journals/applied-mathematics-and-statistics/articles/10.3389/fams.2023.1122114/full.

Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 87(2):362–388, 2025. doi: 10.1093/jrsssb/qkae095. URL https://doi.org/10.1093/jrsssb/qkae095.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/2018/hash/566f0ea4f6c2e947f36795c8f58ba901-Abstract.html.

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, February 2019. doi: 10.1073/pnas.1804597116. URL https://doi.org/10.1073/pnas.1804597116.

Quinn Lanners, Cynthia Rudin, Alexander Volfovsky, and Harsh Parikh. Data fusion for partial identification of causal effects, 2025. URL https://arxiv.org/abs/2505.24296.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Data fusion for efficiency gain in ate estimation: A practical review with simulations, 2024. URL https://arxiv.org/abs/2407.01186.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Combining experimental and observational data through a power likelihood. *Biometrics*, 81(1):ujaf008, 2025. doi: 10.1093/biomtc/ujaf008. URL https://doi.org/10.1093/biomtc/ujaf008.

National Academies of Sciences, Engineering, and Medicine. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. The National Academies Press, Washington, DC, 2022. ISBN 978-0-309-27820-1. doi: 10.17226/26479. URL https://doi.org/10.17226/26479.

Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8398–8407. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/saito20a.html.

Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023. doi: 10.1002/wics.1581. URL https://doi.org/10.1002/wics.1581.

J.A. Vargas-Guzman and A.W. Warrick. Geostatistics for natural resources evaluation. *Journal of Environmental Quality*, 28(3):1044–1044, 1999. doi: 10.2134/jeq1999.00472425002800030046x. URL https://doi.org/10.2134/jeq1999.00472425002800030046x.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1319839.

Sam Witty, Kenta Takatsu, David Jensen, and Vikash Mansinghka. Causal Inference Using Gaussian Processes with Structured Latent Confounders. *Proceedings of Machine Learning Research*, 119:10313–10323, 2020. URL https://proceedings.mlr.press/v119/witty20a.html.

Lili Wu and Shu Yang. Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177, pages 904–926. PMLR, 2022. URL https://proceedings.mlr.press/v177/wu22a.html.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(3):575–596, 2023. doi: 10.1093/jrsssb/qkad017. URL https://doi.org/10.1093/jrsssb/qkad017.

Shu Yang, Margaret Gamalo, and Haoda Fu. Integrating RCTs, RWD, AI/ML and Statistics: Next-Generation Evidence Synthesis, 2025a. URL https://arxiv.org/abs/2511.19735.

Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of treatment effect and confounding function. *Bernoulli*, 31(4):2987 – 3012, 2025b. doi: 10.3150/24-BEJ1835. URL https://doi.org/10.3150/24-BEJ1835.

Appendix A. Identification of CATE in the RCT

Below we show that the CATE can be identified using observed data in the RCT.

$$
\begin{aligned}
\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau] &= \mathbb{E}\big[\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau \cup \mathbf{X}_S] \mid \mathbf{X}_\tau\big] \\
&\quad \text{(tower property)} \\
&= \mathbb{E}\big[\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau\big] \\
&\quad \text{(Assumptions A4)} \\
&= \mathbb{E}\big[\mathbb{E}[Y^1 \mid A = 1, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] - \mathbb{E}[Y^0 \mid A = 0, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau\big] \\
&\quad \text{(Assumptions A2)} \\
&= \mathbb{E}\big[\mathbb{E}[Y^1 \mid A = 1, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau\big] - \mathbb{E}\big[\mathbb{E}[Y^0 \mid A = 0, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau\big] \\
&\quad \text{(linearity of conditional expectation)} \\
&= \mathbb{E}[\mathbb{E}[Y \mid A = 1, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, \mathbf{X}_\tau \cup \mathbf{X}_S, S = \mathrm{e}] \mid \mathbf{X}_\tau] \\
&\quad \text{(Assumption A1)} \\
&= \mathbb{E}[Y \mid A = 1, \mathbf{X}_\tau, S = \mathrm{e}] - \mathbb{E}[Y \mid A = 0, \mathbf{X}_\tau, S = \mathrm{e}] \\
&\quad \text{(tower property and } P(X_S|X_\tau) = P(X_S|X_\tau, S = e)\text{)}
\end{aligned}
$$

Appendix B. Discussion of Assumptions

Assumption A4 is weaker than requiring the RCT to be a random sample from the target population: conditional on $\mathbf{X}_\tau$, it only requires the relevant distribution of selection variables to agree between the trial and the target population. Note that this is strictly weaker than assuming that the RCT is a random sample from the target population and instead allows for some level of distributional shift. Formally, this means that the CATE in the target population can be identified as

$$\tau(X_\tau) = \mathbb{E}[Y^1 - Y^0|X_\tau] = \mathbb{E}_{X_S|X_\tau}[\mathbb{E}[Y^1 - Y^0|X_\tau, X_S, S = e]]$$

Thus averaging over X_S given X_τ in the RCT correctly recovers the CATE. Importantly, this does not require the marginal distribution of X_S to be identical across studies, and distribution shift on X_τ (marginally) is allowed.

Although the CATE can be identified within the RCT population under assumptions A1-A5, this highlights a fundamental limitation of relying on a single data source. The identified estimand, $\tau(X_\tau) = \omega^{(e)}(X_\tau)$, pertains only to the RCT population and therefore reflects the causal effect among individuals who were eligible for and participated in the trial. In many applied settings, however, the primary scientific or policy objective is to infer treatment effects in a broader target population, which may be better represented by an observational study. Assumption A6 formalises the conditions under which such transport is possible: within levels of X_τ, the distribution of X_S is invariant between the RCT and the target population, while still allowing for marginal differences in X_S across the two populations.

Appendix C. Illustrative example

Figure 3 depicts the differences between single independent GPs and multi-task GPs. Independent GPs excel in regions with observed data but exhibit suboptimal performance during extrapolation. In contrast, Multitask GPs enhance both prediction accuracy and uncertainty quantification in unobserved areas, showcasing superior performance in extrapolation scenarios.

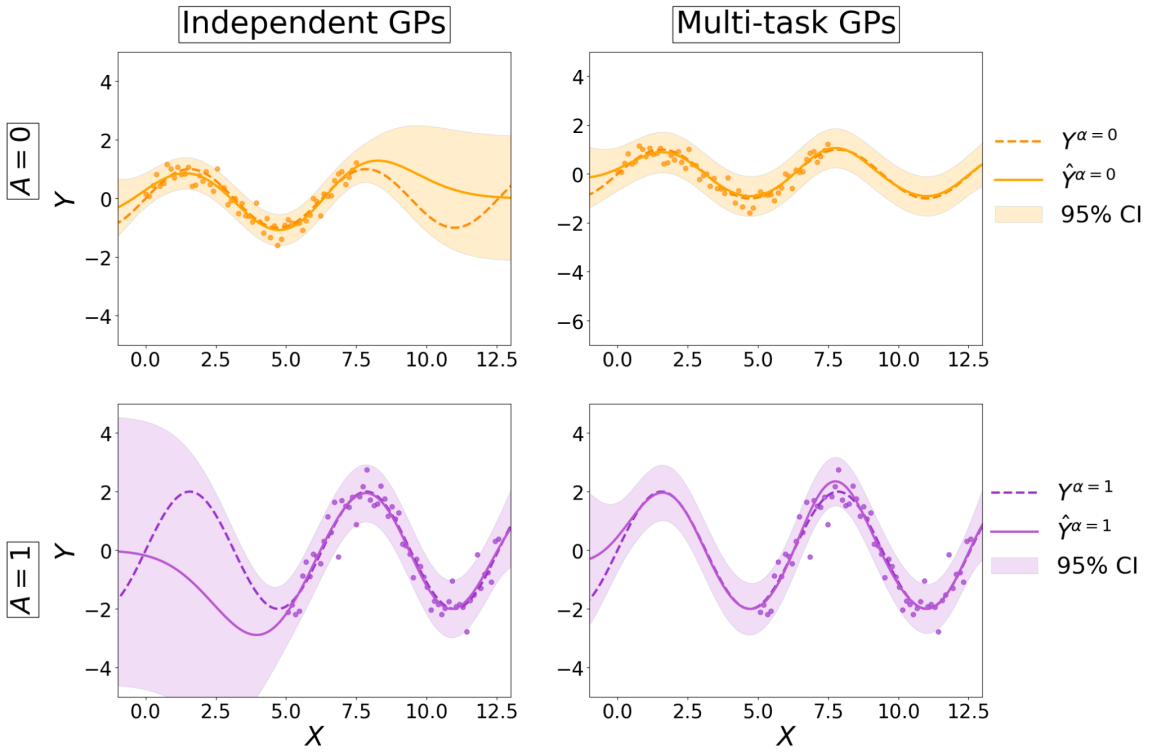

Figure 3: Comparative illustration between Independent GPs and Multitask GP. Here, A represents distinct treatment groups, Y corresponds to the outcome, and X denotes baseline covariates.

Appendix D. Derivation of Posterior Distribution

The ICM set-up is

$$f^{\mathrm{e}}(\mathbf{x}) = \alpha_1^{\mathrm{e}} u_1(\mathbf{x}) + \alpha_2^{\mathrm{e}} u_2(\mathbf{x})$$
$$f^{\mathrm{o}}(\mathbf{x}) = \alpha_1^{\mathrm{o}} u_1(\mathbf{x}) + \alpha_2^{\mathrm{o}} u_2(\mathbf{x})$$

where $u_i(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}))$ independently. Suppose we observe $(\mathbf{y}^{\mathrm{e}}, X^{\mathrm{e}})$ and $(\mathbf{y}^{\mathrm{o}}, X^{\mathrm{o}})$, where the dataset is of size n_{e} and n_{o} respectively. The observational model is

$$\begin{aligned} y_i^{\mathrm{e}} &= f^{\mathrm{e}}(\mathbf{x}_i) + \epsilon_i^{\mathrm{e}} \\ y_i^{\mathrm{o}} &= f^{\mathrm{o}}(\mathbf{x}_i) + \epsilon_i^{\mathrm{o}} \end{aligned}$$

where both $\epsilon^{\mathrm{e}}, \epsilon^{\mathrm{o}}$ arise independently from $\mathcal{N}(0, \sigma^2)$.

We clearly have $\mathbb{E}[y_i^{\mathrm{e}}] = \mathbb{E}[y_i^{\mathrm{o}}] = 0$, where the expectation is over both the observation noise and the GP prior. The variance is more interesting. It is not difficult to show that

$$\mathbb{E}\left[y_i^{\mathrm{e}} y_j^{\mathrm{e}}\right] = \begin{cases} \beta^{\mathrm{e}} k(\mathbf{x}_i^{\mathrm{e}}, \mathbf{x}_j^{\mathrm{e}}) + \sigma^2 & \text{if } i = j \\ \beta^{\mathrm{e}} k(\mathbf{x}_i^{\mathrm{e}}, \mathbf{x}_j^{\mathrm{e}}) & \text{otherwise} \end{cases}$$

Similarly, we have

$$\mathbb{E}\left[y_i^{\mathrm{o}} y_j^{\mathrm{o}}\right] = \begin{cases} \beta^{\mathrm{o}} k(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_j^{\mathrm{o}}) + \sigma^2 & \text{if } i = j \\ \beta^{\mathrm{o}} k(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_j^{\mathrm{o}}) & \text{otherwise} \end{cases}$$

Finally, we have

$$\mathbb{E}\left[y_i^{\mathrm{e}} y_j^{\mathrm{o}}\right] = \beta^{\mathrm{eo}} k(\mathbf{x}_i^{\mathrm{e}}, \mathbf{x}_j^{\mathrm{o}}).$$

Consider a new test point $\mathbf{x}_*$, and we are interested in $f^{\mathrm{e}}(\mathbf{x}_*)$ and $f^{\mathrm{o}}(\mathbf{x}_*)$. One can also show that $\mathbb{E}[f^{\mathrm{e}}(\mathbf{x}_*)] = \mathbb{E}[f^{\mathrm{o}}(\mathbf{x}_*)] = 0$, and

$$\begin{aligned} \mathbb{E}[f^{\mathrm{e}}(\mathbf{x}_*)\, y_i^{\mathrm{e}}] &= \beta^{\mathrm{e}} k(\mathbf{x}_i^{\mathrm{e}}, \mathbf{x}_*) \\ \mathbb{E}[f^{\mathrm{o}}(\mathbf{x}_*) y_i^{\mathrm{o}}] &= \beta^{\mathrm{o}} k(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_*) \\ \mathbb{E}[f^{\mathrm{e}}(\mathbf{x}_*) y_i^{\mathrm{o}}] &= \beta^{\mathrm{eo}} k(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_*) \\ \mathbb{E}[f^{\mathrm{o}}(\mathbf{x}_*) y_i^{\mathrm{e}}] &= \beta^{\mathrm{eo}} k(\mathbf{x}_i^{\mathrm{e}}, \mathbf{x}_*) \end{aligned}$$

We now write this in vector form. Let us define

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^{\mathrm{e}} \\ \mathbf{y}^{\mathrm{o}} \end{bmatrix}$$

which is of length $n_{\mathrm{e}} + n_{\mathrm{o}}$. We can similarly define $X = (X^{\mathrm{e}}, X^{\mathrm{o}})$ which is a $(n_{\mathrm{e}} + n_{\mathrm{o}}) \times p$ matrix. We then write

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, K(X, X) + \sigma^2 I\right)$$

where

$$K(X, X) = \begin{bmatrix} \beta^{\mathrm{e}} K(X^{\mathrm{e}}, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} K(X^{\mathrm{e}}, X^{\mathrm{o}}) \\ \beta^{\mathrm{eo}} K(X^{\mathrm{o}}, X^{\mathrm{e}}) & \beta^{\mathrm{o}} K(X^{\mathrm{o}}, X^{\mathrm{o}}) \end{bmatrix}$$

If we similarly write $\mathbf{f}(\mathbf{x}_*) = (f^{\mathrm{e}}(\mathbf{x}_*), f^{\mathrm{o}}(\mathbf{x}_*))$, then we have

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X, \mathbf{x}_*) \\ K(\mathbf{x}_*, X) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

where

$$\begin{aligned} K(\mathbf{x}_*, \mathbf{x}_*) &= B \otimes k(\mathbf{x}_*, \mathbf{x}_*) \\ K(X, \mathbf{x}_*) &= \begin{bmatrix} \beta^{\mathrm{e}} k(\mathbf{x}_*, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{e}}) \\ \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{o}}) & \beta^{\mathrm{o}} k(\mathbf{x}_*, X^{\mathrm{o}}). \end{bmatrix} \end{aligned}$$

To clarify, $K(X, \mathbf{x}_*)$ is a $(n^{\mathrm{o}} + n^{\mathrm{e}}) \times 2$ matrix.

One can then use the usual conditional of a Gaussian distribution, which shows that

$$\mathbf{f}(\mathbf{x}_*) \mid X, \mathbf{y} \sim \mathcal{N}\left(m_{\mathcal{D}}\left(\mathbf{x}_*\right), K_{\mathcal{D}}\left(\mathbf{x}_*, \mathbf{x}_*\right)\right)$$

where

$$K_{\mathcal{D}}\left(\mathbf{x}_*, \mathbf{x}_*\right) = K\left(\mathbf{x}_*, \mathbf{x}_*\right) - K(\mathbf{x}_*, X)\left[K(X,X) + \sigma^2 I\right]^{-1} K(X, \mathbf{x}_*).$$

The posterior mean can similarly be defined as

$$\mathbf{m}_{\mathcal{D}}\left(\mathbf{x}_*\right) = K(\mathbf{x}_*, X)\left[K(X,X) + \sigma^2 I\right]^{-1} \mathbf{y}.$$

Appendix E. Interpreting ρ

For further intuition, ρ can be interpreted as a measure of codependence between f^{e} and f^{o}, as formalised by the following result.

Proposition 2 *We have $\rho = 1$ if and only if*

$$\alpha_1^{\mathrm{e}} \alpha_2^{\mathrm{o}} = \alpha_1^{\mathrm{o}} \alpha_2^{\mathrm{e}}$$

Proof It is easier to work with $1 - \rho^2$, where plugging in the values from (Equation 1) gives

$$1 - \rho^2 = \frac{\beta^{\mathrm{e}} \beta^{\mathrm{o}} - \left(\beta^{\mathrm{eo}}\right)^2}{\beta^{\mathrm{e}} \beta^{\mathrm{o}}} = \frac{\left(\alpha_1^{\mathrm{e}} \alpha_2^{\mathrm{o}} - \alpha_1^{\mathrm{o}} \alpha_2^{\mathrm{e}}\right)^2}{\beta^{\mathrm{e}} \beta^{\mathrm{o}}}$$

The denominator is positive and finite, so the numerator is thus 0 if and only if $\alpha_1^{\mathrm{e}} \alpha_2^{\mathrm{o}} = \alpha_1^{\mathrm{o}} \alpha_2^{\mathrm{e}}$. ■

Assuming non-zero coefficients, this is equivalent to the condition $\alpha_1^{\mathrm{e}} / \alpha_1^{\mathrm{o}} = \alpha_2^{\mathrm{e}} / \alpha_2^{\mathrm{o}}$, which implies that f^{e} is a scalar multiple of f^{o}. It is thus intuitive that this results in $\rho = 1$, as learning about f^{o} is equivalent to learning about f^{e}. Finally, another useful observation is that if a single coefficient is 0, we will have $\rho < 1$ as long as all other coefficients are non-zero.

Appendix F. Proof of Proposition 3.1

Consider first updating the GP with the experimental data points. This gives us

$$\mathbf{f}(\mathbf{x}_*) \mid X^{\mathrm{e}}, \mathbf{y}^{\mathrm{e}} \sim \mathcal{N}\left(m_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*), K_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*, \mathbf{x}_*)\right)$$

where

$$K_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*, \mathbf{x}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - \begin{bmatrix}\beta^{\mathrm{e}} k(\mathbf{x}_*, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{e}})\end{bmatrix}^T \Sigma_{\mathrm{e}}^{-1} \begin{bmatrix}\beta^{\mathrm{e}} k(\mathbf{x}_*, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{e}})\end{bmatrix}$$

where

$$\Sigma_{\mathrm{e}} = \beta^{\mathrm{e}} K(X^{\mathrm{e}}, X^{\mathrm{e}}) + \sigma^2 I$$

and $K(\mathbf{x}_*, \mathbf{x}_*) = B \otimes k(\mathbf{x}_*, \mathbf{x}_*)$. Note that $\begin{bmatrix}\beta^{\mathrm{e}} k(\mathbf{x}_*, X^{\mathrm{e}}) & \beta^{\mathrm{eo}} k(\mathbf{x}_*, X^{\mathrm{e}})\end{bmatrix}$ is a $n^{\mathrm{e}} \times 2$ matrix. The posterior mean can similarly be defined, but is not our focus here.

We can now simply treat $m_{\mathcal{D}_{\mathrm{e}}}$ and $K_{\mathcal{D}_{\mathrm{e}}}$ as our prior mean and covariance functions respectively, noting that the covariance function can also be written as

$$K_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*, \mathbf{x}_*) = B \otimes k(\mathbf{x}_*, \mathbf{x}_*) - B' \otimes k'(\mathbf{x}_*, \mathbf{x}_*),$$

where

$$B' = \begin{bmatrix}(\beta^{\mathrm{e}})^2 & \beta^{\mathrm{e}}\beta^{\mathrm{eo}} \\ \beta^{\mathrm{e}}\beta^{\mathrm{eo}} & (\beta^{\mathrm{eo}})^2\end{bmatrix}, \qquad k'(\mathbf{x}_*, \mathbf{x}_*) = k(\mathbf{x}_*, X^{\mathrm{e}})^T {\Sigma_{\mathrm{e}}}^{-1} k(\mathbf{x}_*, X^{\mathrm{e}})$$

As in the main paper, let us assume that

$$\begin{bmatrix}\alpha_1^{\mathrm{e}} & \alpha_2^{\mathrm{e}} \\ \alpha_1^{\mathrm{o}} & \alpha_2^{\mathrm{o}}\end{bmatrix} = \begin{bmatrix}1 & 0 \\ \rho & \sqrt{1-\rho^2}\end{bmatrix},$$

which gives

$$B = \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}, \quad B' = \begin{bmatrix}1 & \rho \\ \rho & \rho^2\end{bmatrix}.$$

If we now observe $(X^{\mathrm{o}}, \mathbf{y}^{\mathrm{o}})$, then we can write the full posterior as

$$\mathbf{f}(\mathbf{x}_*) \mid X^{\mathrm{e}}, \mathbf{y}^{\mathrm{e}}, X^{\mathrm{o}}, \mathbf{y}^{\mathrm{o}} \sim \mathcal{N}\left(m_{\mathcal{D}}(\mathbf{x}_*), K_{\mathcal{D}}(\mathbf{x}_*, \mathbf{x}_*)\right).$$

where the key is that

$$K_{\mathcal{D}}(\mathbf{x}_*, \mathbf{x}_*) = K_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*, \mathbf{x}_*) - K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}(\mathbf{x}_*, X^{\mathrm{o}}) \left[K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}(X^{\mathrm{o}}, X^{\mathrm{o}}) + \sigma^2 I\right]^{-1} K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}(X^{\mathrm{o}}, \mathbf{x}_*)$$

where

$$\begin{aligned} K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}(X^{\mathrm{o}}, X^{\mathrm{o}}) &= \beta^{\mathrm{o}} K(X^{\mathrm{o}}, X^{\mathrm{o}}) - (\beta^{\mathrm{eo}})^2 K'(X^{\mathrm{o}}, X^{\mathrm{o}}) \\ &= K(X^{\mathrm{o}}, X^{\mathrm{o}}) - \rho^2 K'(X^{\mathrm{o}}, X^{\mathrm{o}}) \end{aligned}$$

and

$$\begin{aligned} K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}\left(X^{\mathrm{o}}, \mathbf{x}^*\right) &= \left[\beta^{\mathrm{eo}} k(\mathbf{x}^*, X^{\mathrm{o}}) - \beta^{\mathrm{e}}\beta^{\mathrm{eo}} k'(\mathbf{x}^*, X^{\mathrm{o}}), \quad \beta^{\mathrm{o}} k(\mathbf{x}^*, X^{\mathrm{o}}) - (\beta^{\mathrm{eo}})^2 k'(\mathbf{x}^*, X^{\mathrm{o}})\right] \\ &= \left[\rho\left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right), \quad k(\mathbf{x}^*, X^{\mathrm{o}}) - \rho^2 k'(\mathbf{x}^*, X^{\mathrm{o}})\right] \end{aligned}$$

where $K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}\left(X^{\mathrm{o}}, \mathbf{x}^*\right)$ is a matrix of shape $n^{\mathrm{o}} \times 2$ and $K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}\left(\mathbf{x}^*, X^{\mathrm{o}}\right) = K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}\left(X^{\mathrm{o}}, \mathbf{x}^*\right)^T$.

We thus have the posterior variance of f^{e} as

$$\begin{aligned} V^{\mathrm{e}}(\mathbf{x}_*) = &\left(k(\mathbf{x}_*, \mathbf{x}_*) - k'(\mathbf{x}_*, \mathbf{x}_*)\right) \\ &- \rho^2 \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)^T \Sigma_{\mathrm{eo}}^{-1} \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right) \end{aligned}$$

where

$$\begin{aligned} \Sigma_{\mathrm{eo}} &= K^{\mathrm{o}}_{\mathcal{D}_{\mathrm{e}}}\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right) + \sigma^2 I \\ &= K\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right) - \rho^2 K'\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right) + \sigma^2 I. \end{aligned}$$

The first term is the original posterior covariance given $(X^{\mathrm{e}}, \mathbf{y}^{\mathrm{e}})$, whilst the second term is the reduction in variance due to $\mathcal{D}_{\mathrm{o}} = (X^{\mathrm{o}}, \mathbf{y}^{\mathrm{o}})$, which we want to control. In other words, we want to upper bound

$$\rho^2 \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)^T \Sigma_{\mathrm{eo}}^{-1} \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)$$

We now want to write the above term as the posterior variance of a GP. We can guess the following solution and write:

$$V^{\mathrm{e}}(\mathbf{x}_*) = (1-\rho^2)\left(k(\mathbf{x}_*, \mathbf{x}_*) - k'(\mathbf{x}_*, \mathbf{x}_*)\right) + \rho^2 P$$

where

$$\begin{aligned} P = &\left(k(\mathbf{x}_*, \mathbf{x}_*) - k'(\mathbf{x}_*, \mathbf{x}_*)\right) \\ &- \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)^T \Sigma_{\mathrm{eo}}^{-1} \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right). \end{aligned}$$

If we can show that $P \geq 0$, then we are done. To see this, note that we can apply Woodbury's matrix identity which gives

$$\begin{aligned} \Sigma_{\mathrm{eo}}^{-1} &= [\ \underbrace{(1-\rho^2) K'\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right)}_{A} + \underbrace{K\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right) - K'\left(X^{\mathrm{o}}, X^{\mathrm{o}}\right) + \sigma^2 I}_{B}\ ]^{-1} \\ &= B^{-1} - \underbrace{(B + BA^{-1}B)^{-1}}_{C} \end{aligned}$$

which gives

$$\begin{aligned} P = &\left(k(\mathbf{x}_*, \mathbf{x}_*) - k'(\mathbf{x}_*, \mathbf{x}_*)\right) - \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)^T B^{-1} \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right) \\ &+ \underbrace{\left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)^T C \left(k(\mathbf{x}^*, X^{\mathrm{o}}) - k'(\mathbf{x}^*, X^{\mathrm{o}})\right)}_{\geq 0} \end{aligned}$$

To show the final term is positive, we just need to show that C is positive semi-definite which implies $x^T Cx \geq 0$ for any vectors x. If A and B are positive definite (PD) and symmetric, then so is A^{-1} and $BA^{-1}B$. The sum of two PD matrices is PD, as is the inverse of a PD matrix, so C is PD. Finally, we see that the remaining first terms in P is simply the regular posterior variance (given $\mathcal{D}_\text{o}$) of a GP with kernel $k(\mathbf{x}_*, \mathbf{x}_*) - k'(\mathbf{x}_*, \mathbf{x}_*)$, which is non-negative.

Putting this together then, we have

$$V^\text{e}(\mathbf{x}_*) \geq (1-\rho^2)\, V^\text{e}_{\mathcal{D}_\text{e}}(\mathbf{x}_*)$$

where $V^\text{e}_{\mathcal{D}_\text{e}}$ is the variance conditional on $\mathcal{D}_\text{e} = (X^\text{e}, \mathbf{y}^\text{e})$ only.

Appendix G. Confounding function

The hidden confounding function $\eta(\mathbf{x}) = f^\text{e}(\mathbf{x}) - f^\text{o}(\mathbf{x})$ quantifies the degree to which conditional average treatment effect in the observational study deviates from the conditional average treatment effect in the trial. This may be of independent interest in order to understand the underlying process for treatment assignment or allocation in the real-world outside of the experimental setting. Although our principal focus in the above exposition is on f^e, we may also easily obtain a posterior distribution for $\eta(\mathbf{x})$, by considering the joint posterior distribution over $\mathbf{f}(\mathbf{x})$.

G.1. Variance of confounding function

Suppose we now want to compute the variance of the confounding function evaluated at $\mathbf{x}_*$. This is given by $\eta(\mathbf{x}_*) = f^e(\mathbf{x}_*) - f^o(\mathbf{x}_*)$, or in matrix notation $\eta(\mathbf{x}_*) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}^T \mathbf{f}(\mathbf{x}_*)$. By standard properties of the multivariate Gaussian distribution, it follows that

$$\eta(\mathbf{x}_*) \mid X, \mathbf{y} \sim \mathcal{N}\left(m^\text{e}(\mathbf{x}_*) - m^o(\mathbf{x}_*), V_\eta(\mathbf{x}_*)\right),$$

where

$$\begin{aligned}
V_\eta(\mathbf{x}_*) &= \begin{pmatrix} 1 \\ -1 \end{pmatrix}^T K_*(\mathbf{x}_*, \mathbf{x}_*) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\
&= (\beta^\text{e} + \beta^o - 2\beta^{eo}) k(\mathbf{x}_*, \mathbf{x}_*) \\
&\quad - \begin{pmatrix} (\beta^\text{e} - \beta^{eo}) k(\mathbf{x}_*, X^\text{e}) \\ (\beta^{eo} - \beta^o) k(\mathbf{x}_*, X^o) \end{pmatrix}^T \left[K(X,X) + \sigma^2 I\right]^{-1} \begin{pmatrix} (\beta^\text{e} - \beta^{eo}) k(\mathbf{x}_*, X^\text{e}) \\ (\beta^{eo} - \beta^o) k(\mathbf{x}_*, X^o) \end{pmatrix} \\
&= 2(1-\rho) k(\mathbf{x}_*, X^\text{e}) - \begin{pmatrix} (1-\rho) k(\mathbf{x}_*, X^\text{e}) \\ (\rho - 1) k(\mathbf{x}_*, X^o) \end{pmatrix}^T \left[K(X,X) + \sigma^2 I\right]^{-1} \begin{pmatrix} (1-\rho) k(\mathbf{x}_*, X^\text{e}) \\ (\rho - 1) k(\mathbf{x}_*, X^o) \end{pmatrix}
\end{aligned}$$

Appendix H. Multivariate simulation studies

Similar to the univariate case, in the multivariate case we vary the complexity of the confounding and the CATE function, but we choose to explore only cases with nonlinear potential outcomes functions.

Simulation 1: Nonlinear Potential Outcomes, linear confounding, linear CATE

Assume that we have five continuous baseline covariate $X_j \sim U[-2,2]$, where $j = 1, \cdots, 5$. The values of X define the trial participation mechanism defined as $S \sim \text{Bernoulli}(p_S)$ where $p_S = \frac{\exp(-10-8X_1-8X_2)}{1+\exp(-10-8X_1-8X_2)}$. After selecting all trial participants we assign them randomly to treatment $A \sim \text{Bernoulli}(0.5)$. The potential outcomes are generated as $Y(a) = a\tau(X) + \sum_{j=1}^{5} X_j + \epsilon$, for $a = 0, 1$, where $\tau(X) = 1 + X_1 + X_2$ is the CATE and $\epsilon \sim \mathcal{N}(0,1)$. For the observational study we have $X_j \sim U[-2,2]$, where $j = 1, \cdots, 5$, and the treatment is generated according to the model $A \sim \text{Bernoulli}(e(X))$ where $\text{logit}(e(X)) = -(X_1 + X_2)$. Similarly to the trial the potential outcomes are $Y(a) = a\tau(X) + \sum_{j=1}^{5} X_j + U + \epsilon$, for $a = 0, 1$. U represents the hidden confounding and we generated as $U \sim \mathcal{N}((2A-1)(X_1 + X_2), 1)$.

Simulation 2: Nonlinear Potential Outcomes, nonlinear confounding, nonlinear CATE

Assume that we have five continuous baseline covariate $X_j \sim U[-2,2]$, where $j = 1, \cdots, 5$. The values of X define the trial participation mechanism defined as $S \sim \text{Bernoulli}(p_S)$ where $p_S = \frac{\exp(-10-8X_1-8X_2)}{1+\exp(-10-8X_1-8X_2)}$. After selecting all trial participants we assign them randomly to treatment $A \sim \text{Bernoulli}(0.5)$. The potential outcomes are generated as $Y(a) = a\tau(X) + \sum_{j=1}^{5} X_j + \epsilon$, for $a = 0, 1$, where $\tau(X) = 1 + X_1 + X_1^2 + X_2 + X_2^2$ is the CATE and $\epsilon \sim \mathcal{N}(0,1)$. For the observational study we have $X_j \sim U[-2,2]$, where $j = 1, \cdots, 5$, and the treatment is generated according to the model $A \sim \text{Bernoulli}(e(X))$ where $\text{logit}(e(X)) = -(X_1 + X_2)$. Similarly to the trial the potential outcomes are $Y(a) = a\tau(X) + \sum_{j=1}^{5} X_j + U + \epsilon$, for $a = 0, 1$. U represents the hidden confounding and we generated as $U \sim \mathcal{N}((2A-1)(\sin(X_1) + \sin(X_2)), 1)$.

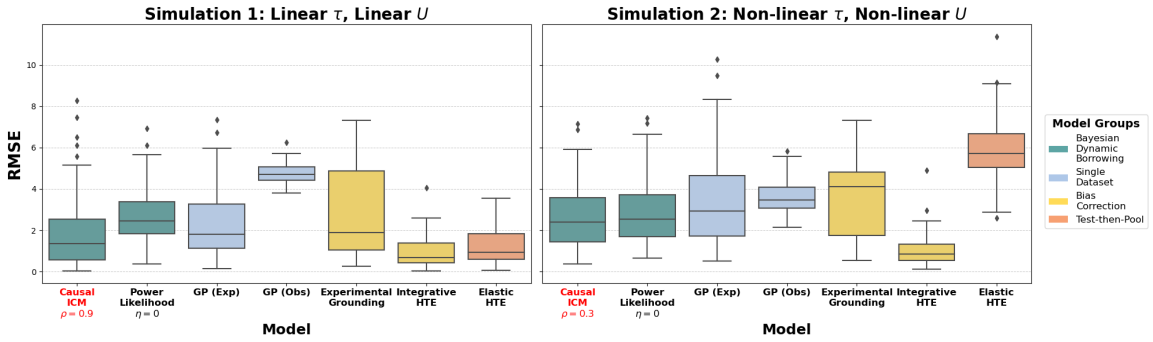

Figure 4: Multivariate simulation results over 100 simulated datasets. *Left*: Simulation setting 1: linear CATE and unobserved confounding, *Right*: Non-linear CATE and unobserved confounding

Based on Figure 4, in the first simulation setting, the highest performance is achieved by the Integrative HTE and the Elastic HTE, as expected given that both models are correctly specified. Notably, the performance of *Causal-ICM* is comparable to these benchmarks, demonstrating its competitiveness even when the models are well-specified.

In the second simulation setting, where both the CATE and the confounding exhibit more complex, non-linear structures, the performance of models relying on correct specification deteriorates. In contrast, *Causal-ICM* maintains robust performance, being outperformed

only by the Integrative HTE. Importantly, this superior performance of the Integrative HTE is a consequence of having a correctly specified outcome model; In the multivariate simulation settings, parametric competitors can outperform Causal-ICM when their outcome and confounding models are correctly specified. Table 2 shows, however, that under misspecification Causal-ICM becomes more robust and substantially outperforms Integrative HTE in the second multivariate simulation setting.

Table 2: RMSE in the second multivariate simulation setting under model misspecification.

Model	Mean RMSE (SD)
GP (exp)	3.422 (2.186)
GP (obs)	3.580 (0.796)
Experimental Grounding	3.582 (1.713)
Integrative HTE	21.883 (3.330)
Causal-ICM	2.586 (1.513)

Appendix I. Extrapolation with *Causal-ICM*

Below we show the performance of *Causal-ICM* in the first univariate simulation setting, where both the CATE and the confounding functions are non-linear. The *Causal-ICM* can successfully produce accurate estimates of the CATE both within and outside the support of the RCT with inflated uncertainty outside the support.

Appendix J. Sensitivity to ρ, Observational Sample Size, Kernel Choice, and Degree of Overlap; and Performance Within and Outside the RCT Support

Below we present a more thorough exploration of the effects of different values of ρ, differing observational/experimental sample sizes, the choice of kernel and the degree of overlap between the two studies for the second simulation setting of the main text where the CATE and the confounding function are non-linear.

ρ	**RMSE**
0.0	2.571
0.2	2.051
0.4	1.502
0.6	0.913
0.8	0.305
1.0	1.095

Table 3: RMSE values for different values of ρ, Univariate simulation 2

The results above highlight the importance of ρ in controlling the influence of unobserved confounding: as ρ increases from 0 to 0.8, the RMSE decreases steadily, reflecting a stronger

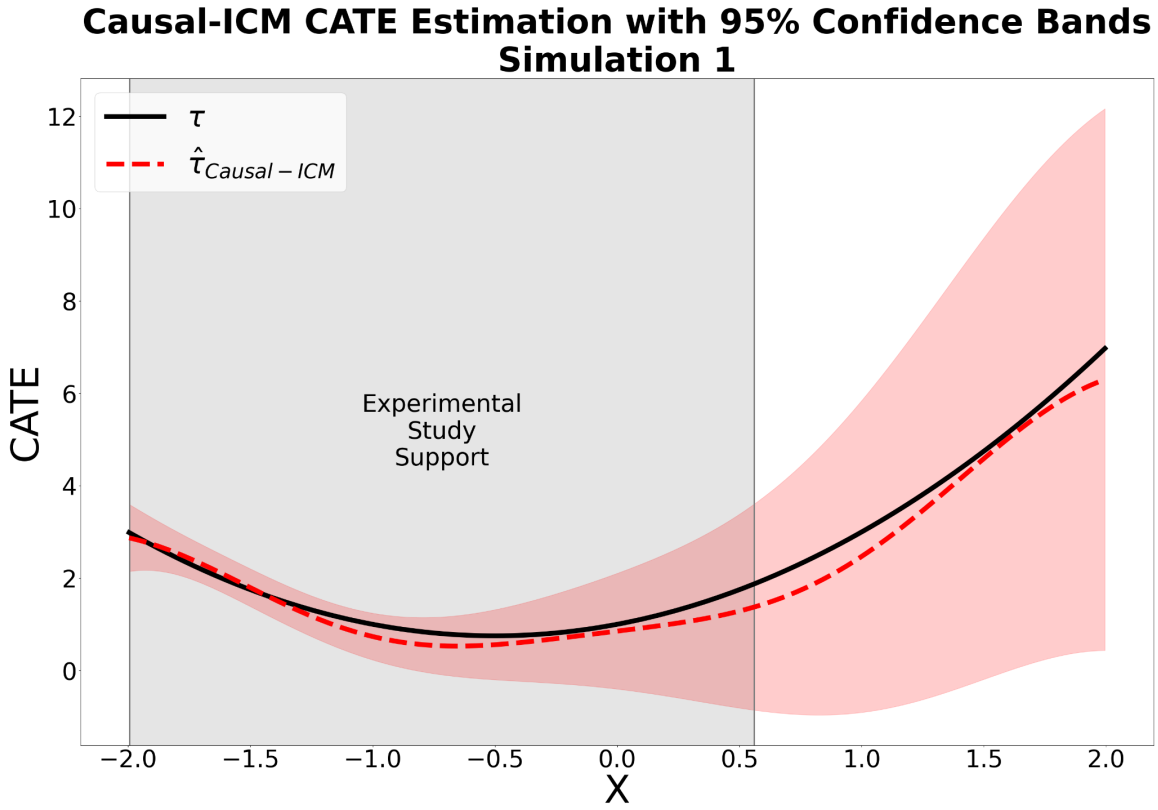

Figure 5: Estimated CATE under the first univariate simulation setting with nonlinear CATE and confounding functions.

Kernel	Mean RMSE (SD)
RBF	0.822 (0.651)
Matérn 3/2	1.315 (0.0556)
Matérn 5/2	0.968 (0.583)

Table 4: Mean RMSE for different kernels, Univariate simulation 2

borrowing of information from the observational study, before rising again at $\rho = 1.0$ where the confounding is no longer accounted for. Across different kernel choices, *Causal-ICM* retains strong and consistent performance, with the RBF kernel achieving the lowest mean RMSE, suggesting some robustness to this modelling choice, though the differences are moderate.

Table 5 reports results under increasing observational sample size while keeping the experimental sample size fixed. *Causal-ICM* remains competitive as the observational sample grows substantially larger than the RCT, and in fact improves with larger observational samples, in contrast to *Integrative HTE* whose performance degrades noticeably. The performance of the experimental grounding method of Kallus et al. (2018), the power

Table 5: Performance (RMSE) under increasing observational sample size in simulation setting 2. Results are based on 50 simulated datasets.

Model	Observational sample size	Mean RMSE (SD)
Causal-ICM (ours)	200	1.078 (0.810)
Integrative HTE	200	1.080 (0.441)
Power Likelihood	200	1.023 (0.559)
Experimental Grounding	200	1.691 (0.597)
GP Observational	200	1.512 (0.226)
Causal-ICM (ours)	500	1.071 (0.652)
Integrative HTE	500	1.406 (0.461)
Power Likelihood	500	1.231 (0.667)
Experimental Grounding	500	1.646 (0.547)
GP Observational	500	1.475 (0.146)
Causal-ICM (ours)	1000	0.929 (0.695)
Integrative HTE	1000	1.546 (0.303)
Power Likelihood	1000	0.975 (0.506)
Experimental Grounding	1000	1.652 (0.452)
GP Observational	1000	1.414 (0.128)
Causal-ICM (ours)	2000	0.934 (0.674)
Integrative HTE	2000	1.576 (0.318)
Power Likelihood	2000	1.161 (0.606)
Experimental Grounding	2000	1.660 (0.443)
GP Observational	2000	1.377 (0.062)

Overlap	Optimal ρ	**Mean RMSE (SD)**
Full	0.5	0.264 (0.088)
High	0.4	0.444 (0.237)
Low	0.8	0.941 (0.607)

Table 6: Mean RMSE for different degrees of overlap between the RCT and observational study for the second simulation setting of the main text (Univariate simulation 2); The degree of overlap is controlled through the coefficients of the model for study participation; Results obtained over 100 different simulated datasets.

likelihood of Lin et al. (2025), and the T-learner trained on observational data does not seem to deteriorate with the increasing observational sample size, but it remains inferior to our method. Finally, Table 6 shows that *Causal-ICM* performs well even under low overlap between the two studies, with optimal results obtained under full overlap, as expected. Taken together, these results suggest that *Causal-ICM* is robust across a range of practical conditions that may arise when combining experimental and observational data.

To complement the aggregate RMSE results in the main text, Table 7 reports mean squared error separately inside and outside the support of the RCT for simulation setting 2.

Table 7: Mean in-sample and out-of-sample MSE in simulation setting 2. In-sample refers to the support of the randomized trial; out-of-sample refers to regions outside the trial support.

Method	Mean in-sample MSE (SD)	Mean out-of-sample MSE (SD)
GP exp	15.302 (15.595)	9.645 (13.259)
GP obs	343.387 (436.318)	310.823 (412.965)
Kallus GP	3.048 (3.141)	43.461 (42.069)
Integrative HTE	0.307 (0.114)	3.153 (0.919)
Causal-ICM	0.252 (0.099)	1.071 (0.722)

Causal-ICM achieves the lowest in-sample MSE among all methods, and most strikingly, retains strong performance outside the RCT support, where competing methods deteriorate considerably. In particular, *Kallus GP* degrades sharply out-of-sample despite performing reasonably well within the trial support, suggesting it does not generalise well beyond the experimental region. *Causal-ICM*, by contrast, benefits from the broader covariate coverage of the observational study to extrapolate more reliably, underscoring one of the key practical advantages of integrating the two data sources.

Appendix K. Runtime results

Training multi-task Gaussian processes scales cubically in the total sample size, although the intrinsic coregionalization structure can often be exploited to improve efficiency in practice. For larger datasets, sparse GP approximations with inducing points provide a natural route to scalability.

Table 8: Runtime performance in seconds for simulation setting 2, based on 100 datasets.

Method	Mean (s)	SD (s)	Median (s)	Min (s)	Max (s)
GP exp	0.244369	0.114056	0.236458	0.089087	0.489595
GP obs	2.232287	0.639261	2.077287	1.417782	6.207515
Experimental Grounding GP	2.394518	0.951331	2.181754	1.566869	10.213725
Causal-ICM	4.559152	0.782487	4.513439	3.310538	8.185129
Integrative HTE	6.006100	0.416126	5.964500	5.222000	7.494000