B-CALM: Bias-Limited Bayesian Borrowing for RCT-Defined Conditional Average Treatment Effects under Covariate Mismatch

Amir Asiaee*
Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN 37203, USA

Samhita Pal
Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN 37203, USA

Abstract

Randomized controlled trials (RCTs) identify treatment effects in the randomized trial population but are often too small for reliable heterogeneity estimation. Observational studies (OS) are larger but confounded and measured on only partially overlapping covariates. We develop Bayesian Calibrated ALignment under covariate Mismatch (B-CALM), a Bayesian borrowing method for RCT-defined conditional average treatment effect (CATE) estimation. B-CALM maps source-specific covariates into a shared latent state, jointly models trial and observational outcome surfaces, and uses baseline-bias and comparative-bias functions to represent how the OS departs from the trial estimand. The comparative-bias prior acts as a sensitivity parameter. In a linearized model with a fixed representation, we prove a bias-limited information bound showing that observational contrast information about the trial treatment-effect function is capped by the prior precision of this bias function, and derive an effective-sample-size formula showing that the RCT-equivalent information contributed by the OS saturates as OS sample size grows. The theory also combines a PAC-Bayes trial-risk bound with an integral-probability-metric (IPM) alignment and calibration decomposition that separates RCT empirical risk, latent alignment, and residual calibration of the debiased OS surface. In synthetic, semi-synthetic, and pediatric-obesity external-control studies, B-CALM maintains near-nominal average coverage and low negative transfer while pooled and causal-forest baselines become overconfident under comparative bias.

1 Introduction

Randomized controlled trials (RCTs) remain the main design for internally valid treatment-effect estimation. In many clinical and policy settings, however, RCTs are powered for an average treatment effect, not for conditional average treatment effects (CATEs) over patient subgroups. Large observational studies (OS) offer the obvious source of additional information. They contain many more individuals, often measured through detailed health systems, registries, or administrative databases. That information carries its own bias. Treatment assignment is not randomized, unmeasured confounding can distort treatment contrasts, patient populations can differ from the trial, and the measured covariates in the two sources often overlap only partially.

The problem is therefore not simply how to combine an RCT and an OS. The harder task is to let the OS reduce uncertainty about the *trial-population* CATE while preventing a large but biased source from becoming the de facto estimand. Recent calibrated-borrowing methods were built around this constraint. R-OSCAR uses observational outcome predictions as nuisance functions and calibrates their residual discrepancy in the trial [5], MR-OSCAR extends the same calibration strategy by imputing trial-missing observational covariates [30], and CALM avoids explicit imputation by aligning source-specific embeddings before calibration [2]. These methods use the RCT to define

*Corresponding author: amir.asiaeetaheri@vumc.org.

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

and identify the target contrast, but primarily return point estimates. Bayesian borrowing takes a different route. Power priors, commensurate priors, robust meta-analytic-predictive priors, and joint external-control models use priors or explicit bias parameters to regulate external evidence [9, 17, 20, 31], but usually target scalar or low-dimensional treatment summaries, such as average effects, response rates, or subgroup effects, rather than a covariate-indexed CATE function.

Modeling principle. Bayesian Calibrated ALignment under covariate Mismatch (B-CALM) views the observational study as a biased, partially mismatched measurement of the trial outcome surfaces. Let trial covariates be $X^r = (W, U)$ and observational covariates be $X^o = (W, V)$, where W is shared, U is trial-only, and V is observational-only. Source-specific encoders map each covariate block into a common latent patient state Z. The latent state is not itself the estimand. It is the coordinate system for comparing the two data sources at matched patient states.

On this shared coordinate system, the RCT defines the trial outcome surface

$$\eta^r(a, z) = \mu(z) + a\tau(z).$$

Here $\mu(z)$ is the trial control-outcome surface, the prognosis at $A = 0$, and $\tau(z)$ is the trial treatment-contrast surface. Thus $\eta^r(0, z) = \mu(z)$ and $\eta^r(1, z) - \eta^r(0, z) = \tau(z)$. Learning τ is the CATE problem after integrating over the trial encoder. Both μ and τ are unknown functions.

The observational study is allowed to have a different surface:

$$\eta^o(a, z) = \mu(z) + a\tau(z) + b_0(z) + ab_\Delta(z).$$

The function $b_0(z)$ is the *baseline bias*, the difference between OS and RCT controls at the same latent patient state. The function $b_\Delta(z)$ is the *comparative bias*, the difference between the OS treatment contrast and the RCT treatment contrast after the baseline shift is removed. Equivalently, the OS contrast is $\tau(z) + b_\Delta(z)$, while the trial contrast of interest remains $\tau(z)$. This decomposition is the causal accounting of B-CALM. The OS may inform prognosis, representation, and even treatment contrasts, but only through an explicit model for how it may be biased relative to the RCT.

Borrowing as sensitivity. Priors on b_0 and b_Δ control borrowing. A tight prior on b_Δ encodes the belief that OS and RCT treatment contrasts are commensurate at each latent state, so the OS can sharpen inference for τ. A diffuse prior allows the OS treatment contrast to be arbitrarily biased, so comparative borrowing reverts toward RCT-only inference. This is a function-valued analogue of commensurate and power-prior borrowing [9, 20, 21]. Instead of discounting the entire external likelihood by one scalar, B-CALM places a prior on the covariate-indexed discrepancy that determines which parts of the OS contrast are allowed to inform the trial contrast.

Main claim. The key consequence is *bias-limited borrowing*. If the comparative-bias prior has nonzero variance, observational information about the trial treatment-effect function has a finite ceiling, no matter how large the OS becomes. In a linearized function model, the OS Fisher information for τ is bounded by the prior precision of b_Δ. In the scalar special case for a one-number treatment summary, the effective sample size (ESS) saturates at a level determined by the ratio of observational noise to prior comparative-bias variance. This makes sensitivity analysis concrete. Changing the prior standard deviation for comparative bias changes the maximum amount of treatment-contrast information the OS is allowed to contribute.

Contributions and positioning. We make four contributions.

1. We develop **Bayesian calibrated alignment** for RCT-defined CATE estimation in the partially overlapping-covariate setting, yielding a posterior over encoders, outcome surfaces, bias functions, and trial-population CATEs.
2. We introduce a **baseline-bias** and **comparative-bias** decomposition that separates the trial estimand τ from the observational contrast $\tau + b_\Delta$ after alignment onto the shared latent state.
3. We prove **function-valued bias-limited borrowing** for a fixed representation and a finite basis. The comparative-bias prior precision bounds the information observational contrasts can contribute about the trial treatment-effect surface, with a scalar ESS corollary for one-number treatment summaries and the limiting full-pooling and RCT-only regimes.
4. We evaluate the closed-form Gaussian instance of the model on synthetic, semi-synthetic, and pediatric-obesity external-control studies. Across the main factorial simulation study, B-CALM has near-nominal 90% credible-interval coverage (0.939 on average, with a minimum of about 0.86 at intermediate comparative bias) with low negative transfer (0.023), while pooled and causal-forest baselines become overconfident under comparative bias. In the real-data study, B-CALM achieves a modest width reduction while keeping sensitivity to external-control bias explicit.

Relative to R-OSCAR, MR-OSCAR, and CALM [2, 5, 30], B-CALM casts calibrated borrowing as posterior sensitivity analysis, rather than a point estimate with bootstrap. The representation and the calibration target are inherited from CALM. What is new is the borrowing calculus placed on them. Appendix D.7 compares B-CALM with commensurate and robust-MAP priors. Relative to power, commensurate, robust MAP, and dynamic borrowing priors [9, 20, 24, 27, 31], B-CALM does not introduce a single global borrowing weight, such as a likelihood discount or exchangeability indicator. The amount of borrowing is implied locally by the prior and posterior uncertainty in the bias function $b_\Delta(z)$, learned jointly with source-specific encoders.

2 Related work

Observational borrowing for trial estimands. The closest line of work starts from observational outcome models to improve trial-based CATE estimation while using the randomized trial to define and identify the RCT-defined CATE. R-OSCAR calibrates observational conditional-mean predictions using RCT data, reducing variance while preserving unbiased trial contrasts under regularity conditions [5]. MR-OSCAR carries the calibration idea to covariate mismatch by imputing observational-only variables in the trial and deriving a finite-sample decomposition with an imputation-error term [30]. CALM, the direct predecessor of B-CALM, replaces explicit imputation with source-specific embeddings aligned into a shared representation and decomposes risk into alignment, outcome-model, and calibration terms [2]. B-CALM inherits this calibrated-alignment target and the same estimand discipline. Observational information can reduce uncertainty only through calibration or explicit bias modeling, not by redefining the RCT estimand toward the OS contrast. The change is inferential rather than estimand-level. B-CALM turns calibrated alignment from a point-estimation procedure with separately propagated uncertainty into a Bayesian joint model, yielding a posterior over encoders, trial and observational outcome surfaces, source-bias functions, and trial-population CATEs. A robustness-based alternative is the QR-learner [26], a model-agnostic learner that uses external trial or observational data to reduce the error of a trial CATE estimate while still recovering the trial CATE when the external data are misaligned, with no prior to elicit. B-CALM instead models the misalignment through an explicit prior. Its protection rests on that prior, which is varied in a sensitivity analysis, and in exchange it returns a posterior for the trial CATE. Appendix D.8 compares the two on the same simulated designs.

Bayesian borrowing and external controls. B-CALM's borrowing mechanism sits nearer to Bayesian external-data methods, which regulate how much historical or external evidence enters an analysis. Power priors downweight the historical likelihood by a scalar [9], commensurate priors borrow according to parameter similarity [20], robust meta-analytic-predictive priors add a weakly informative mixture component to protect against prior-data conflict [29, 31], and multisource exchangeability models use discrete indicators to decide which sources are exchangeable [24], with heterogeneous-effect dynamic borrowing studied in Kotalik et al. [27]. These methods answer the question of *how much* external evidence to use, typically for average effects or matched-covariate models. Reviews by van Rosmalen et al. [35] and Lewis et al. [28] emphasize the danger of uncontrolled prior-data conflict. B-CALM asks the borrowing question pointwise in latent state. How much should the OS treatment contrast at state z inform the posterior for the RCT treatment contrast at the same state? The prior on $b_\Delta(z)$ answers that question. Tight priors permit local pooling, while diffuse priors return the contrast toward RCT-only inference. This is a function-valued version of the explicit bias parameters advocated in joint external-control modeling [17].

Representation learning and RCT–OS fusion. Representation-based CATE methods address treatment-arm or domain mismatch within one study [8, 23, 32], and Bayesian CATE learners and RCT–OS fusion methods [1, 16, 18, 22, 25] model CATE uncertainty or correct observational models with trial data. Closest to our Bayesian framing, Causal-ICM [12] uses multi-task Gaussian processes with a borrowing parameter, but it does not target the RCT-defined CATE under partially overlapping covariates or make source bias an explicit baseline and comparative process. Appendix G discusses these lines in more detail.

3 Problem setup

We observe two samples: a randomized trial $\mathcal{D}^r = \{(Y_i^r, A_i^r, W_i^r, U_i^r) : i = 1, \dots, n^r\}$ and an observational study $\mathcal{D}^o = \{(Y_j^o, A_j^o, W_j^o, V_j^o) : j = 1, \dots, n^o\}$. The treatment $A \in \{0, 1\}$ is randomized in $\mathcal{D}^r$ but not in $\mathcal{D}^o$. The covariate block W is shared, U is trial-only, and V is

observational-only. We use $s \in \{r, o\}$ for source. The source indicator is written as a superscript throughout (n^r, X^r, η^o, R^r). The only exception is the residual variances σ_r^2, σ_o^2.

Let $Y^r(a)$ denote the trial potential outcome under treatment a. We focus on continuous outcomes. The trial estimand is

$$\tau^r(x^r) = \mathbb{E}\{Y^r(1) - Y^r(0) \mid X^r = x^r\}. \tag{1}$$

Binary outcomes can be handled by placing the same bias decomposition on the linear-predictor scale of a generalized linear model and transforming posterior draws back to risks and risk differences. Appendix A.8 provides the notation. The OS superscript denotes a data source, not a second causal estimand. Because A^o is not randomized, OS treatment-level outcome surfaces are treated as observed conditional surfaces that may be biased relative to the trial causal surface.

Assumption 1 (Trial identification)**.** *Within the RCT, treatment is randomized and satisfies positivity:* $0 < \mathbb{P}(A^r = 1 \mid X^r) < 1$. *Consistency holds:* $Y^r = Y^r(A^r)$.

Assumption 2 (Trial-relevant latent state)**.** *Write* $m_a^r(x^r) = \mathbb{E}\{Y^r(a) \mid X^r = x^r\}$. *There exist a trial encoder* $q_0^r(z \mid x^r)$, *functions* $\mu_0, \tau_0 : \mathcal{Z} \to \mathbb{R}$, *and a representation error* $\varepsilon_{\text{rep}} \geq 0$ *such that, for* $a \in \{0, 1\}$,

$$\sup_{x^r} \left| m_a^r(x^r) - \mathbb{E}_{Z \sim q_0^r(\cdot \mid x^r)}\{\mu_0(Z) + a\tau_0(Z)\} \right| \leq \varepsilon_{\text{rep}}. \tag{2}$$

This assumption says that the latent state is a sufficient working coordinate system for the RCT outcome surfaces, up to approximation error. It is not an assumption that Z is a literal biological hidden variable. If the encoder is a deterministic map $z = e^r(x^r)$, Eq. (2) reduces to $m_a^r(x^r) \approx \mu_0\{e^r(x^r)\} + a\tau_0\{e^r(x^r)\}$.

In the theory, we use an integral probability metric (IPM) to quantify the remaining difference between the induced trial and OS latent distributions. For probability distributions P_1 and P_2 on $\mathcal{Z}$ and a class of test functions $\mathcal{G}$, define:

$$\text{IPM}_{\mathcal{G}}(P_1, P_2) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{Z \sim P_1}\{g(Z)\} - \mathbb{E}_{Z \sim P_2}\{g(Z)\} \right|. \tag{3}$$

Maximum mean discrepancy is one special case, obtained when $\mathcal{G}$ is the unit ball of a reproducing-kernel Hilbert space [14]. The assumption below records the residual source mismatch after the chosen encoder construction or regularization. It applies to any encoder construction that satisfies the stated IPM bound.

Assumption 3 (Calibratable source alignment)**.** *Let* $P_Z^r(\phi^r)$ *and* $P_Z^o(\phi^o)$ *denote the latent distributions induced by* $Z^r \sim q_{\phi^r}(\cdot \mid X^r)$ *and* $Z^o \sim q_{\phi^o}(\cdot \mid X^o)$. *For a class of test functions* $\mathcal{G}$, *define*

$$\Delta_Z(\phi^r, \phi^o) = \text{IPM}_{\mathcal{G}}\{P_Z^r(\phi^r), P_Z^o(\phi^o)\}. \tag{4}$$

There exist source-specific encoders (ϕ_0^r, ϕ_0^o), *with* ϕ_0^r *satisfying Assumption 2, such that* $\Delta_Z(\phi_0^r, \phi_0^o) \leq \varepsilon_{\text{align}}$ *for some residual alignment error* $\varepsilon_{\text{align}}$.

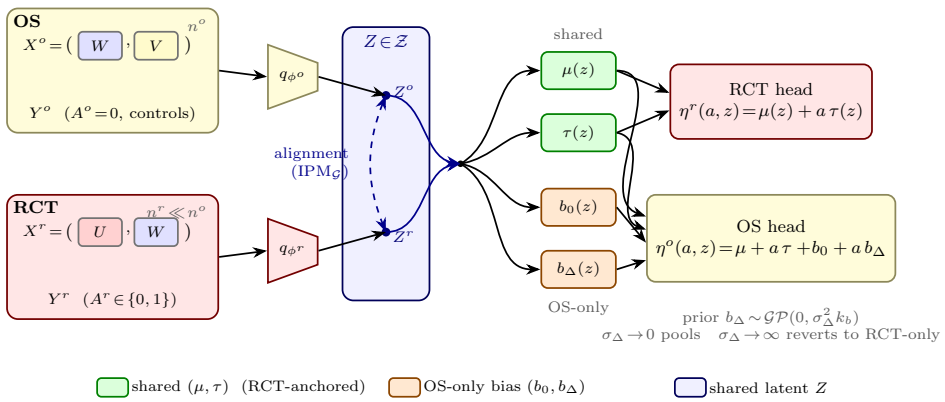


Figure 1: B-CALM architecture and causal bookkeeping. Source-specific encoders q_{ϕ^r} and q_{ϕ^o} map the partially overlapping covariate blocks $X^r = (W, U)$ and $X^o = (W, V)$ into a shared latent state Z, with source alignment encoded by the chosen latent maps and, when those maps are learned, regularized by an IPM or related calibration loss. The RCT head defines the RCT-defined functions (μ, τ), while the OS head observes the shifted functions $(\mu + b_0, \tau + b_\Delta)$. Priors on the bias functions, especially b_Δ, determine how much OS contrast information can enter the posterior for the trial CATE.

We make no ignorability assumption for the observational treatment assignment. Instead, the observational conditional outcome surface is allowed to differ from the trial causal surface through bias functions. This separates B-CALM from point-identification designs that presume observational treatment comparisons are causal. The observational data are useful but not presumed causal for treatment comparisons.

4 B-CALM model

Figure 1 summarizes our framework. B-CALM introduces source-specific encoders $Z_i^r \sim q_{\phi^r}(z \mid W_i^r, U_i^r)$ and $Z_j^o \sim q_{\phi^o}(z \mid W_j^o, V_j^o)$, where q_{ϕ^r} and q_{ϕ^o} may be deterministic neural encoders, variational encoders, sparse linear projections, or GP latent-variable maps. The corresponding source-specific latent distributions are the distributions of Z_i^r when trial covariates are passed through q_{ϕ^r} and of Z_j^o when OS covariates are passed through q_{ϕ^o}. The encoder does not make the OS causal. It defines a common coordinate system on which trial and OS outcome surfaces can be compared, while the remaining source discrepancy is absorbed by explicit bias functions. In the general framework, learned encoders can be regularized by matching latent distributions, for example through the IPM in Eq. (3). Shared-covariate reconstruction can be added if one specifies an auxiliary decoder.

We define the outcome model on the shared latent space. For our continuous-outcome model, the likelihood is Gaussian. Write $p_{\mathrm{G}}(y; m, \sigma^2)$ for the $\mathcal{N}(m, \sigma^2)$ density. The RCT likelihood is

$$Y_i^r \mid A_i^r, Z_i^r, \mu, \tau \sim \mathcal{N}\{\eta^r(A_i^r, Z_i^r), \sigma_r^2\}, \quad \eta^r(a, z) = \mu(z) + a\tau(z). \tag{5}$$

The OS likelihood has the same Gaussian form but with source-bias functions:

$$Y_j^o \mid A_j^o, Z_j^o, \mu, \tau, b_0, b_\Delta \sim \mathcal{N}\{\eta^o(A_j^o, Z_j^o), \sigma_o^2\}, \quad \eta^o(a, z) = \mu(z) + a\tau(z) + b_0(z) + ab_\Delta(z). \tag{6}$$

The function μ is the trial baseline outcome surface and τ is the trial treatment-effect function. The two source-specific bias functions record how the OS departs from the trial. Evaluating Eq. (6) at $a = 0$ and $a = 1$, the control-surface gap $\eta^o(0, z) - \eta^r(0, z) = b_0(z)$ is the **baseline bias** (how OS controls differ from RCT controls at the same latent z, through unobserved comorbidities, selection, or measurement drift) and the contrast gap $[\eta^o(1, z) - \eta^o(0, z)] - [\eta^r(1, z) - \eta^r(0, z)] = b_\Delta(z)$ is the **comparative bias** (the residual that confounding leaves on the contrast after the baseline shift is removed). The OS has its own pair $(\mu + b_0, \tau + b_\Delta)$ for the control surface and treatment effect, while the RCT identifies (μ, τ). The posterior CATE is computed by integrating the encoder:

$$\tau^r(x^r) = \mathbb{E}\left[\tau(Z) \mid Z \sim q_{\phi^r}(\cdot \mid x^r), \mathcal{D}\right], \tag{7}$$

Priors and borrowing interpretation. One nonparametric specification places independent GP priors $\mu \sim \mathcal{GP}(0, k_\mu)$, $\tau \sim \mathcal{GP}(0, k_\tau)$, $b_0 \sim \mathcal{GP}(0, \sigma_0^2 k_b)$, and $b_\Delta \sim \mathcal{GP}(m_\Delta, \sigma_\Delta^2 k_b)$. The mean m_Δ can encode directional prior knowledge about confounding. The variance σ_Δ^2 controls comparative borrowing. If $\sigma_\Delta^2 = 0$ and $m_\Delta = 0$, the observational and randomized treatment contrasts are pooled directly. If $\sigma_\Delta^2 \to \infty$, the OS treatment contrast is allowed to differ arbitrarily from the RCT contrast and therefore contributes essentially no comparative information about τ. Intermediate values define a sensitivity analysis.

This prior standard deviation connects B-CALM to classical Bayesian borrowing. A power prior discounts an external likelihood globally. A commensurate prior borrows according to similarity between low-dimensional parameters. B-CALM instead places the commensurability question on the contrast discrepancy $b_\Delta(z)$ itself. The OS can be highly precise about its own contrast $\tau(z) + b_\Delta(z)$, but its information about the trial contrast $\tau(z)$ is limited by how tightly the analyst is willing to concentrate b_Δ around its prior mean.

Computation. With learned encoders, the posterior over basis or network weights and encoder parameters is approximated variationally (Appendix B.3). All experiments in Section 6 use the closed-form Gaussian instance with fixed encoders, where no approximation is needed.

5 Theory

This section studies how much the OS can say about the trial contrast and when that information carries over to the trial population. Assumption 1 fixes the RCT estimand, so the OS can sharpen it but not redefine it, and Assumption 2 lets us write the trial surfaces on Z, where the bias functions and their priors are also defined. We proceed in four steps. First, we consider the two extremes of the comparative-bias prior. With $b_\Delta \equiv 0$ the OS contrast is pooled directly, and as the prior on b_Δ becomes flat the OS contributes no information about τ (Proposition 1, Appendix A.2). Second, we reduce the estimand to a single number, such as an ATE, and derive the effective sample size behind

bias-limited borrowing. Third, we repeat the calculation for a finite-basis representation of the CATE, where the prior precision of b_Δ bounds what the OS can say about the whole treatment-effect surface. These three results hold for a fixed representation and a linear Gaussian contrast model, and Section 6 examines how the fitted method behaves. Finally, we bring in Assumption 3 through Δ_Z and state a risk decomposition for general posteriors that separates RCT empirical risk, latent alignment, and residual calibration of the debiased OS surface.

Effective sample size for an ATE. For intuition and comparison with classical borrowing, first collapse the estimand to a one-number treatment-effect summary, such as the trial-population ATE $\bar{\tau} = \mathbb{E}_{X^r}\{\tau^r(X^r)\}$. This is not the main B-CALM estimand. It is the finite-dimensional special case where effective-sample-size language is most transparent. The same calculation applies to any fixed finite-dimensional summary of the CATE surface, such as a prespecified subgroup treatment effect.

Corollary 1 (ATE special case: bias-limited effective sample size)**.** *Take the one-dimensional special case $\widehat{\bar{\tau}}^r \mid \bar{\tau} \sim \mathcal{N}(\bar{\tau}, \sigma_r^2/n^r)$ and $\widehat{\bar{\tau}}^o \mid \bar{\tau}, b \sim \mathcal{N}(\bar{\tau} + b, \sigma_o^2/n^o)$, with priors $b \sim \mathcal{N}(0, \sigma_b^2)$ and $\bar{\tau} \sim \mathcal{N}(m_0, v_0)$, where $\bar{\tau}$ is a scalar, one-number treatment contrast and m_0, v_0 are its prior mean and prior variance. Then the posterior precision for $\bar{\tau}$ is $V_n^{-1} = v_0^{-1} + n^r/\sigma_r^2 + (\sigma_o^2/n^o + \sigma_b^2)^{-1}$, the OS contributes the same precision as an unbiased RCT contrast with effective sample size $n^o_{\text{eff}} = n^o/(1 + n^o\sigma_b^2/\sigma_o^2)$, and $\lim_{n^o\to\infty} n^o_{\text{eff}} = \sigma_o^2/\sigma_b^2$.*

Increasing n^o can remove sampling noise in the OS contrast, but it cannot remove prior uncertainty about the OS-to-RCT contrast gap. The saturation level is determined by the ratio of observational noise to comparative-bias prior variance.

Function-valued bias-limited information. The one-number calculation becomes a B-CALM statement when applied to a latent treatment-effect surface. Here the encoder and function basis play different roles. The encoder parameters ϕ determine where a unit is placed in latent space. They answer the question, "what is this unit's latent state z?" The basis map ψ is used after z has been chosen. It answers the question, "how do we represent functions such as $\tau(z)$ and $b_\Delta(z)$ on that latent space?" The matrix Φ^o then stacks the OS basis rows $\psi(Z_j^o)^\top$ and is therefore a design matrix for the OS contrast regression, not an encoder.

Theorem 1 (Function-valued bias-limited information, fixed representation)**.** *Fix the encoders and hence the OS latent locations $Z_1^o, \ldots, Z_{n^o}^o$. Let $\psi(z) \in \mathbb{R}^d$ be a basis-function map, write $\tau(z) = \psi(z)^\top\beta_\tau$ and $b_\Delta(z) = \psi(z)^\top\beta_b$, and assume $\beta_b \sim \mathcal{N}(0, \Sigma_b)$ with Σ_b positive definite. Stack the OS basis rows $\psi(Z_j^o)^\top$ into $\Phi^o \in \mathbb{R}^{n^o\times d}$. Suppose the OS contrast data are summarized by the Gaussian linear model $\widehat{d}^o = \Phi^o\beta_\tau + \Phi^o\beta_b + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n^o})$, where $\widehat{d}^o$ is a noisy vector of OS treatment-contrast measurements at the OS latent locations. Marginalizing over β_b gives $\widehat{d}^o \mid \beta_\tau \sim \mathcal{N}\{\Phi^o\beta_\tau, \sigma^2 I_{n^o} + \Phi^o\Sigma_b(\Phi^o)^\top\}$. If Φ^o has full column rank, the OS Fisher information for β_τ is*

$$I^o = (\Phi^o)^\top\left\{\sigma^2 I + \Phi^o\Sigma_b(\Phi^o)^\top\right\}^{-1}\Phi^o = \left\{\Sigma_b + \sigma^2((\Phi^o)^\top\Phi^o)^{-1}\right\}^{-1}, \quad I^o \preceq \Sigma_b^{-1}. \tag{8}$$

The positive-semidefinite statement means that the OS information is no larger in any finite-dimensional direction than the comparative-bias prior precision. As OS information grows in every basis direction, for example when the smallest eigenvalue of $(\Phi^o)^\top\Phi^o$ diverges, $I^o \to \Sigma_b^{-1}$.

The algebra above is standard Gaussian marginalization. The B-CALM implication is the important part. Once the OS contrast is modeled as $\tau(z) + b_\Delta(z)$, the prior precision of the unknown function b_Δ is the information ceiling for borrowing about the unknown function τ. Appendix A.4 extends the calculation to uncertain encoders. Encoder uncertainty adds to the comparative-bias variance in the borrowing cap.

Calibrated-alignment risk decomposition. The information bounds above explain the borrowing cap. They do not by themselves guarantee that a learned representation transfers trial risk well. The final result combines a PAC-Bayes bound for RCT empirical risk with a deterministic IPM-calibration decomposition for the debiased OS surface. Write ρ for a posterior over full model draws $f = (\phi^r, \phi^o, \mu, \tau, b_0, b_\Delta)$ and ρ_0 for a data-independent prior over f. Let $\eta_f^r(a, z) = \mu_f(z) + a\tau_f(z)$ be the RCT surface under draw f, let $\eta_\star^r$ and $\eta_\star^o$ be the true RCT and OS conditional outcome surfaces on the latent space, and define the debiased OS mean under draw f by $\eta^o_{\text{deb},f}(a, z) = \eta_\star^o(a, z) - b_{0,f}(z) - ab_{\Delta,f}(z)$. Let P_m denote the Gaussian outcome distribution indexed by mean m, define $\mathcal{L}_\ell(t, m) = \mathbb{E}_{Y\sim P_m}\ell(t, Y)$, and let $\pi^r(a \mid z)$ be the trial treatment distribution used to define the risk. For a fixed latent state z, set $r_f^r(z) = \sum_{a=0}^1 \pi^r(a \mid z)\mathcal{L}_\ell\{\eta_f^r(a, z), \eta_\star^r(a, z)\}$ and $\widetilde{r}_f^o(z) =$

$\sum_{a=0}^{1} \pi^r(a \mid z)\mathcal{L}_\ell\{\eta_f^r(a,z), \eta^o_{\mathrm{deb},f}(a,z)\}$. Also let $\widehat{L}^r(f) = (n^r)^{-1}\sum_{i=1}^{n^r} \ell\{\eta_f^r(A_i^r, Z_i^r), Y_i^r\}$. Then

$$R^r(\rho) = \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P_Z^r(f)} r_f^r(Z), \quad \widehat{R}^r(\rho) = \mathbb{E}_{f\sim\rho}\widehat{L}^r(f), \quad \widetilde{R}^o(\rho) = \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P_Z^o(f)}\widetilde{r}_f^o(Z). \quad (9)$$

Here $R^r(\rho)$ is the trial-population risk, $\widehat{R}^r(\rho)$ is its empirical RCT analogue, and $\widetilde{R}^o(\rho)$ is the OS risk after subtracting the modeled baseline and comparative biases. To define the calibration residual, let $d_\ell(u,v) = \sup_t |\mathcal{L}_\ell(t,u) - \mathcal{L}_\ell(t,v)|$, and set

$$\varepsilon_{\mathrm{cal}}(\rho) = \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P_Z^r(f)} \sum_{a=0}^{1} \pi^r(a \mid Z)\, d_\ell\{\eta_\star^r(a,Z), \eta^o_{\mathrm{deb},f}(a,Z)\} \quad (10)$$

The residual $\varepsilon_{\mathrm{cal}}(\rho)$ vanishes when the debiased OS surface matches the RCT surface on the trial latent support.

Theorem 2 (PAC-Bayes and IPM-calibration target-risk decomposition)**.** *Assume the loss is bounded in* $[0,1]$ *and that the IPM regularity condition stated in Appendix A.6 holds for* ρ*. Informally, the debiased OS conditional risk must be controlled by the class of test functions* $\mathcal{G}$ *up to a constant* L*. Define* $\Delta_Z(\rho) = \mathbb{E}_{f\sim\rho}\,\mathrm{IPM}_{\mathcal{G}}\{P_Z^r(f), P_Z^o(f)\}$*. With the quantities in Eqs. (9)–(10), for any data-independent prior* ρ_0 *over* f*, with probability at least* $1-\delta$ *over the RCT sample, every such posterior* ρ *satisfies*

$$R^r(\rho) \le \min\left[\widehat{R}^r(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\rho_0) + \log(2\sqrt{n^r}/\delta)}{2n^r}},\ \widetilde{R}^o(\rho) + L\Delta_Z(\rho) + \varepsilon_{\mathrm{cal}}(\rho)\right]. \quad (11)$$

The first term inside the minimum is the RCT-only PAC-Bayes certificate. The second is the OS IPM-calibration certificate. Unlike Theorem 1, the decomposition is not restricted to a fixed representation or a Gaussian model. It holds for every posterior over model draws, learned encoders included, under the bounded-loss and IPM regularity conditions. Its OS term depends on unknown surfaces, so it guides design.

This decomposition explains why alignment alone is insufficient. Small $\Delta_Z(\rho)$ only establishes that the two latent distributions are close. Useful borrowing also requires the debiased OS surface to be calibrated to the trial surface on the shared support, and Theorem 1 limits how much treatment-contrast precision can be taken from the OS while maintaining that protection.

6 Experiments

The experiments evaluate the four claims from Section 1: (i) *bias-limited borrowing* — OS contrast information saturates at the ceiling predicted by Corollary 1, (ii) *negative-transfer protection* — borrowing rarely increases error relative to an RCT-only analysis, (iii) *calibration* — 90% credible intervals stay close to nominal coverage, and (iv) *competitive point accuracy*. Three studies address them: a synthetic factorial study in which the true CATE is known by construction (claims i–iv), a semi-synthetic benchmark modeled on the Tennessee STAR class-size trial (claims iii–iv), and a pediatric-obesity external-control application with real trial and EHR data (claims ii–iii). Appendix B gives the full specifications, and code is at `https://github.com/AsiaeeLab/b-calm`.

6.1 Synthetic factorial study

Design. A three-dimensional latent state generates trial covariates $X^r = (W, U)$ and observational covariates $X^o = (W, V)$ through source-specific noisy linear maps. This keeps the true CATE known even though the sources share only the block W. The factorial grid crosses three covariate-overlap regimes (*shared-only*, *balanced*, and *mismatch-heavy*, with increasingly source-specific covariate loadings and increasingly nonlinear outcome surfaces), three OS bias regimes (none, baseline-only, comparative), and, on the balanced DGP, a sweep that scales the comparative-bias function by $\{0, 0.2, 0.5, 1.0\}$, using $n^r \in \{250, 500\}$, $n^o = 5{,}000$, and 50 replicates per cell. OS treatment assignment is confounded through the observational covariates in every cell. Appendix B.2 gives the generative equations, the assignment model, the full condition grid, and generation pseudocode.

B-CALM configuration. All synthetic experiments use the closed-form Gaussian instance of the model in Section 4. Encoders are fixed identity maps, so each source's covariate vector serves as its latent state. The functions μ, τ, b_0, and b_Δ are represented on a prespecified polynomial–trigonometric basis, and the posterior over basis coefficients is Gaussian in closed form. We compute posterior CATE means, variances, and 90% intervals analytically, without Monte Carlo sampling. No encoder,

decoder, or alignment penalty is trained. The comparative-bias scale is selected within each replicate by empirical-Bayes model averaging over the fixed candidate set $\sigma_\Delta \in \{0.01, 0.05, 0.2, 0.5, 1.0, \infty\}$ with global fixed prior weights that are asymmetric by design (skeptical on the tight scales, a mild mode at moderate borrowing, and a heavy diffuse tail) and prior mean $m_\Delta = 0$. Algorithm 2 in Appendix B.4 states the fitting and selection procedure. These experiments isolate bias-limited borrowing in a fixed common working representation. They do not evaluate learned latent alignment.

Baselines. We compare with RCT-only and OS-only fits of the same Gaussian model, pooled RCT+OS (the same model with $b_0 = b_\Delta \equiv 0$), a causal forest (CF) fit to the pooled data [36], and the calibrated-borrowing family R-OSCAR, MR-OSCAR, and CALM [2, 5, 30]. In our implementation, R-OSCAR builds the OS base surface from the shared covariates alone. MR-OSCAR ridge-imputes the source-specific OS covariates for trial units and fits on both blocks. CALM imputes them the same way and projects both blocks onto a partial-least-squares embedding fit to the OS outcome, a single embedding rather than CALM's source-specific encoders (Appendix A.12). All three then pass their base surface through the same trial-anchored calibration, with the same doubly robust pseudo-outcome and the same CATE regression on the trial design. In these designs the calibration absorbs nearly all of the difference. The three are separate implementations whose results differ in every replicate, but their grid means agree to within 10^{-7}, four orders of magnitude below the Monte Carlo standard error (Appendix D.4). Table 1 and the figures report them jointly as OSCAR/CALM. Bayesian causal forests (BCF) [16] require a full MCMC run per fit, so we evaluate them in the real-data study, where the number of fits is small.

Metrics. We score each replicate on 600 fresh covariate draws from the trial distribution, where the true CATE is known: RMSE of the posterior-mean CATE, empirical coverage and mean width of the pointwise 90% credible intervals, and negative transfer (NT), the fraction of replicates in which a method's RMSE exceeds the RCT-only RMSE on the same replicate. Formulas are in Appendix B.1.

Results. Table 1 summarizes the four metrics over the factorial grid. Per-cell results are in Appendix C. B-CALM is the only borrowing method whose average coverage reaches the nominal 0.90 (0.939), while also improving accuracy, with RMSE 0.206 against 0.318 for RCT-only and near-zero negative transfer (0.023). The two point-accuracy leaders lose calibration. OS-only (RMSE 0.193) covers at 0.677, and pooled has the narrowest intervals (width 0.299) at 0.581 coverage and a 0.292 negative-transfer rate.

Table 1: Synthetic factorial study: averages over the 24-cell grid (50 replicates per cell). Cov. and Width refer to pointwise 90% credible intervals; NT is the negative-transfer rate relative to RCT-only. Per-cell results: Table 3.

Method	RMSE	Cov.	Width	NT
B-CALM	0.206	0.939	0.739	0.023
OSCAR/CALM	0.228	0.890	0.742	0.013
RCT-only	0.318	0.902	1.017	0.000
pooled	0.305	0.581	0.299	0.292
OS-only	0.193	0.677	0.345	0.192
CF	0.308	0.811	0.764	0.299

Figure 2 isolates the comparative-bias axis. The balanced DGP is fixed while the comparative-bias function is scaled up, holding baseline bias and OS confounding constant. The RMSE panel shows the accuracy tradeoff across methods. Pooled and CF are the most accurate when the OS contrast is unbiased, but their errors grow as the bias strengthens (pooled RMSE reaches 1.226 at the strongest setting). OSCAR/CALM is nearly flat across the sweep and beats B-CALM under strong bias (RMSE 0.199 versus 0.239 at strength 0.5), because it declines to borrow contrast information, which protects it there but also means it gains nothing when the OS is clean (0.201 versus B-CALM's 0.167 at strength 0). In the coverage panel, B-CALM stays near nominal at every bias level while pooled falls to 0.07 and CF to 0.23 at the strongest setting. B-CALM's weakest cell is 0.857 coverage at strength 0.5, the transition region where the empirical-Bayes weights still place mass on borrowing scales (the exported posterior weight on the moderate scale $\sigma_\Delta = 0.2$ averages 0.91 there, collapsing to 0.03 by strength 1.0). We return to this cost of adaptivity in the Discussion. Grouped by bias regime (Appendix D.5), B-CALM covers 0.961 without comparative bias, 0.928 at strength 0.2, 0.858 at 0.5, and 0.895 at 1.0, with negative transfer at most 0.10. At strength 0.5 the non-borrowing OSCAR/CALM family is better calibrated (0.892).

Observational sample-size domination. This experiment fixes n^r and grows n^o from 500 to 50,000 under comparative bias, fitting B-CALM at fixed $\sigma_\Delta \in \{0.1, 0.2, 0.4, \infty\}$ with prior mean $m_\Delta = 0.2$ (Appendix A.12). Pooled and OS-only intervals contract as $(n^o)^{-1/2}$ regardless of bias. B-CALM intervals contract until they reach the floor implied by σ_Δ^2 and then saturate (Appendix D.1, Figure 3). At the largest n^o the empirical effective sample size tracks the theoretical saturation value $\sigma_o^2/\sigma_\Delta^2$ (for example, 105.6 empirical versus 99.8 theoretical at $\sigma_\Delta = 0.2$; Appendix D.1,

Table 4). The small nonzero empirical ESS at $\sigma_\Delta = \infty$ (about 4) comes from the finite numerical implementation of the diffuse prior and information flowing through the shared baseline surface, not comparative borrowing.

Ablation. On the balanced DGP at comparative-bias strength 0.2, removing the comparative-bias function collapses coverage from 0.915 to 0.324 while RMSE moves far less (0.195 to 0.259), so b_Δ matters mainly for calibration. Forcing $b_0 \equiv 0$, with the same empirical-Bayes averaging over σ_Δ, also under-covers (0.791, with negative transfer 0.28) (Appendix D.3, Table 6).

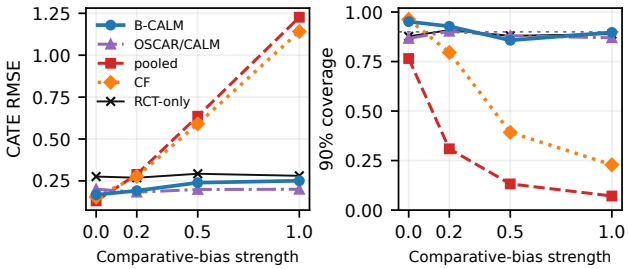


Figure 2: Comparative-bias sweep on the balanced DGP ($n^r = 500$, $n^o = 5{,}000$, 50 replicates per point): CATE RMSE (left) and coverage of pointwise 90% intervals (right; dashed line is nominal). R-OSCAR, MR-OSCAR, and CALM differ by less than 10^{-4} at every point (Appendix D.4) and are drawn as one OSCAR/CALM line.

Posterior calibration curves across nominal levels match the 90% snapshot. B-CALM tracks the diagonal while pooled and CF under-cover (Appendix E, Figure 5).

Decision operating characteristics. We also evaluate the rule "Go if the posterior probability of a positive trial-population ATE is at least 0.975," whose nominal false-Go rate is 0.025, with the effect shifted so that the population ATE is exactly 0 or 0.2 (200 replicates per cell; Appendix D.9). Without comparative bias, B-CALM's false-Go rate is at most 0.010 in every regime and OS size, and its power exceeds that of the trial-only rule (0.73–0.83 versus 0.58–0.62). Under comparative bias the false-Go rate exceeds the nominal level (0.035–0.060 at strength 0.2 and 0.045–0.110 at 0.5, against 0.015–0.035 for trial-only and above 0.85 for pooling). Higher power does not offset this excess. The decision guarantee rests on the comparative-bias prior. In the balanced design, fixing $\sigma_\Delta \geq 0.5$ keeps the false-Go rate at or below 0.035, with power close to trial-only.

Weak alignment. The three covariate regimes weaken the alignment between the sources, though they also add outcome nonlinearity and covariate noise, so they cannot isolate alignment. Along them the trial–OS MMD on the working representation grows from 0.016 to 0.256, B-CALM's intervals widen by 32–38%, and its coverage (0.926–0.966) and negative transfer stay flat, with intervals a quarter to a third narrower than trial-only throughout (Appendix D.6).

6.2 Semi-synthetic benchmark modeled on the STAR trial

The Tennessee Student/Teacher Achievement Ratio (STAR) class-size trial is a standard covariate template for semi-synthetic causal benchmarks [25]. Covariates simulated from parametric marginals modeled on the trial are combined with known, imposed potential-outcome surfaces, so evaluation against the true CATE remains possible. Following the construction of Asiaee et al. [5], we generate a trial-like sample ($n^r = 500$) and an observational sample ($n^o = 5{,}000$) over seven STAR-type covariates (baseline score, sex, race, free-lunch status, small-class assignment, teacher experience, and school quality), with source-shifted marginals that mimic RCT-versus-EHR mismatch, confounded OS treatment assignment, and known baseline and comparative bias functions. The full routine is in Appendix B.5. Because the OS sample contains both arms, with confounded assignment and a comparative bias of known magnitude, this benchmark is where the comparative-bias channel is tested against a known CATE. The real-data application in Section 6.3 exercises only the baseline channel. Over 50 replicates, B-CALM and pooled have similar observed RMSE (0.191 and 0.194; paired difference -0.002, standard error 0.009), while B-CALM raises coverage from pooled's 0.425 to 0.936. MR-OSCAR, the family member run in this study, covers at 0.889 with RMSE 0.228 (Appendix D.2, Table 5). Varying σ_Δ traces posterior CATEs at seven representative covariate profiles that interpolate between pooled ($\sigma_\Delta \to 0$) and RCT-only ($\sigma_\Delta \to \infty$) inference (Appendix D.2, Figure 4). This sensitivity curve is the report a B-CALM analysis would provide.

6.3 Real-data application: external-control augmentation

Data and identification. As a secondary methodological application, we fit B-CALM to data from the Greenlight Plus trial [19], an individually randomized pediatric-obesity prevention trial at six centers. The EHR cohort covers three of the six centers, so the analysis is restricted to complete cases at those three ($n^r = 385$), augmented with a cross-sectional EHR cohort ($n^o = 15{,}150$)

from the same health systems and treated as external controls. The outcome is 12-month weight-for-length z-score. The six shared covariates are site, sex, race/ethnicity, language, insurance, and baseline weight-for-length z-score, with five trial-only covariates retained for sensitivity analyses. Appendix B.6 describes the cohort, preprocessing, and model configuration. Because $A_j^o \equiv 0$ on every EHR row, the OS likelihood involves only $\mu(z) + b_0(z)$. The comparative-bias prior σ_Δ is unidentified and held diffuse, and the baseline-commensurability prior σ_0 on $b_0(z)$ controls borrowing strength. The application therefore evaluates baseline-bias borrowing only.

Sensitivity analysis. Table 2 reports the trial-population ATE across the baseline-bias scale σ_0. Its endpoints are the pooled fit ($\sigma_0 = 0$) and the trial-only fit ($\sigma_0 = \infty$). Over the whole range the estimate moves from -0.189 to -0.171, about 10%, while the interval widens by 40%. The prior acts mainly on the uncertainty, not on the estimate. This flatness is the prior-data-conflict diagnostic here, since the EHR controls do not pull the estimate away from the trial-only value. The 90% interval excludes zero at every evaluated scale, including the unrestricted prior, but its upper limit is within 0.002 of zero for $\sigma_0 \geq 0.5$, so no tipping point is reached on the grid and the sign is resolved only narrowly. A prior-weighted score (Appendix B.6) selects $\sigma_0 = 0.05$, a 10.5% width reduction. We report it as one point on the curve, not as a borrowing strength learned from the data, because b_0 is not identified. Causal forest, BCF [16], and pooling reach much larger reductions by implicitly setting $b_0 \equiv 0$. Appendix F reports these baselines, a bias-susceptibility diagnostic, and leave-one-covariate masking.

Table 2: Real-data ATE across the baseline-bias scale σ_0, with exact 90% intervals. Width reduction is relative to trial-only. The asterisk marks the score-selected scale. All baselines: Table 15.

σ_0	ATE [90% CI]	Width (Red.)
0 (pooled)	$-0.189\,[-0.310, -0.068]$	$0.243\,(28.6\%)$
0.025	$-0.187\,[-0.324, -0.051]$	$0.273\,(19.8\%)$
0.05^*	$-0.184\,[-0.336, -0.032]$	$0.304\,(10.5\%)$
0.1	$-0.178\,[-0.342, -0.015]$	$0.327\,(3.9\%)$
0.2	$-0.174\,[-0.342, -0.006]$	$0.336\,(1.2\%)$
0.5	$-0.171\,[-0.341, -0.002]$	$0.339\,(0.2\%)$
∞ (trial-only)	$-0.171\,[-0.341, -0.001]$	0.340
CF	$-0.162\,[-0.211, -0.112]$	$0.099\,(70.9\%)$

7 Discussion and conclusion

Discussion. The function-valued bias-limited information bound (Theorem 1) and its scalar instance (Corollary 1) give a closed-form prediction that the experiments confirm. Observational borrowing produces a σ_Δ-dependent uncertainty floor instead of asymptotically zero credible-interval width. The same mechanism explains the one weak cell in the bias sweep. At intermediate comparative bias the empirical-Bayes weights retreat from borrowing gradually, and coverage dips to 0.857 before recovering. The same cost of adaptivity appears in decisions. A Go/No-Go rule built on the averaged posterior exceeds its nominal false-positive rate under comparative bias (Section 6). In practice, this lets an analyst report a sensitivity analysis over prespecified values of σ_Δ, rather than committing to one borrowing weight, and compute the n_{eff}^o implied by each value. The pediatric-obesity application illustrates the same pattern with one-arm OS data. Once the baseline-bias prior allows plausible EHR-versus-trial control bias, the width reduction is modest (10.5% at $\sigma_0 = 0.05$ and under 4% for $\sigma_0 \geq 0.1$), while the larger reductions reported by pooled, CF, and BCF rest on the implicit and unverifiable assumption that $b_0 \equiv 0$.

Practical interpretation. The practical output of B-CALM is not a single borrowed estimate, but a calibrated borrowing analysis: a posterior for the RCT-defined estimand, a sensitivity parameter for OS contrast bias, and an ESS diagnostic describing how much information the OS is allowed to contribute. The reported estimand changes accordingly. Rather than asking whether the OS is "similar enough" to pool, the analyst reports how conclusions move as σ_Δ or σ_0 varies, and whether the resulting interval contraction is commensurate with the bias the model is willing to absorb. B-CALM is most useful when external data are potentially informative but not trustworthy enough to be treated as randomized evidence.

Limitations. (i) b_Δ is not identifiable without prior restrictions, which is why we report a sensitivity curve over prespecified σ_Δ values. (ii) Under comparative bias, coverage falls to 0.858 at strength 0.5 and the false-positive rate of the Go/No-Go rule exceeds its nominal level, because the decision guarantee rests on the prior rather than the data. (iii) When the comparative bias is constant in the covariates, a scalar discrepancy is the more efficient estimator (RMSE 0.156 versus 0.215; Appendix D.7). Future directions are in Appendix A.11.

Acknowledgments and Disclosure of Funding

We thank William J. Heerman, Lee M. Sanders, and Jonathan S. Schildcrout for the development and conduct of the Greenlight Plus trial and for the data resource used in Section 6.3, and Jared D. Huling. The Greenlight Plus trial was funded by the Patient-Centered Outcomes Research Institute (PCORI), contract AD-2018C1-11238. A.A. and S.P. were supported by the PCORI award ME-2023C1-32148. The authors declare no competing interests.

References

[1] 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, pages 3424–3432, 2017.

[2] Amir Asiaee and Samhita Pal. Improving RCT-based treatment effect estimation under covariate mismatch via calibrated alignment. *arXiv preprint arXiv:2603.19186*, 2026.

[3] Amir Asiaee, Samet Oymak, Kevin R. Coombes, and Arindam Banerjee. Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *ICML Workshop on Adaptive and Multitask Learning*, 2019.

[4] Amir Asiaee, Samet Oymak, Kevin R. Coombes, and Arindam Banerjee. High dimensional data enrichment: Interpretable, fast, and data-efficient. *arXiv preprint arXiv:1806.04047*, 2023.

[5] Amir Asiaee, Chiara Di Gravio, Cole Beck, Yuting Mei, Samhita Pal, and Jared D. Huling. Improving precision of RCT-based CATE estimation using data borrowing with double calibration. *arXiv preprint arXiv:2306.17478*, 2026.

[6] Amir Asiaee, Samhita Pal, Cole Beck, and Jared D. Huling. Sharp bounds for treatment effect generalization under outcome distribution shift. *arXiv preprint arXiv:2602.09595*, 2026. To appear, 5th Conference on Causal Learning and Reasoning (CLeaR).

[7] Amir Asiaee, Samhita Pal, and Jared D. Huling. Omitted-variable sensitivity analysis for generalizing randomized trials. *arXiv preprint arXiv:2603.27788*, 2026.

[8] Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. In *Advances in Neural Information Processing Systems*, volume 35, pages 37184–37198, 2022.

[9] Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000. doi: 10.1214/ss/1009212673.

[10] 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*, 39(1):165–191, 2024. doi: 10.1214/23-STS889.

[11] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.

[12] Evangelos Dimitriou, Edwin Fong, Jens Magelund Tarp, Karla Diaz-Ordaz, and Brieuc Lehmann. Causal-ICM: A data fusion framework for heterogeneous treatment effect estimation with multi-task Gaussian processes. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pages 248–276, 2026.

[13] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004. doi: 10.1145/1014052.1014067.

[14] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[15] Samuel M. Gross and Robert Tibshirani. Data shared Lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235, 2016. doi: 10.1016/j.csda.2016.02.015.

[16] 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. doi: 10.1214/19-BA1195.

[17] Frank E. Harrell. Incorporating historical control data into an RCT. Statistical Thinking blog, 2023. URL https://www.fharrell.com/post/hxcontrol/. Accessed 2026-04-25.

[18] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

[19] William J. Heerman, Russell L. Rothman, Lee M. Sanders, Jonathan S. Schildcrout, Kori B. Flower, Alan M. Delamater, Melissa C. Kay, Charles T. Wood, Rachel S. Gross, Aihua Bian, Laura E. Adams, Evan C. Sommer, H. Shonna Yin, Eliana M. Perrin, and Greenlight Investigators. A digital health behavior intervention to prevent childhood obesity: The Greenlight Plus randomized clinical trial. *JAMA*, 332(24): 2068–2080, 2024. doi: 10.1001/jama.2024.22362.

[20] Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011. doi: 10.1111/j.1541-0420.2011.01564.x.

[21] Brian P. Hobbs, Daniel J. Sargent, and Bradley P. Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3):639–674, 2012. doi: 10.1214/12-BA722.

[22] Maximilian Ilse, Patrick Forré, Max Welling, and Joris M. Mooij. Combining interventional and observational data using causal reductions. *arXiv preprint arXiv:2103.04786*, 2023.

[23] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, 2016.

[24] Alexander M. Kaizer, Joseph S. Koopmeiners, and Brian P. Hobbs. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics*, 19(2):169–184, 2018. doi: 10.1093/biostatistics/kxx031.

[25] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[26] Rickard K. A. Karlsson, Piersilvio De Bartolomeis, Issa Dahabreh, and Jesse H. Krijthe. Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 1783–1791, 2026.

[27] Ales Kotalik, David M. Vock, Eric C. Donny, Dorothy K. Hatsukami, and Joseph S. Koopmeiners. Dynamic borrowing in the presence of treatment effect heterogeneity. *Biostatistics*, 22(4):789–804, 2021. doi: 10.1093/biostatistics/kxz066.

[28] Connor Jo Lewis, Somnath Sarkar, Jiawen Zhu, and Bradley P. Carlin. Borrowing from historical control data in cancer drug development: A cautionary tale and practical guidelines. *Statistics in Biopharmaceutical Research*, 11(1):67–78, 2019. doi: 10.1080/19466315.2018.1497533.

[29] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J. Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010. doi: 10.1177/1740774509356002.

[30] Samhita Pal, Jared D. Huling, and Amir Asiaee. Improving RCT-based CATE estimation under covariate mismatch via double calibration. *arXiv preprint arXiv:2603.17066*, 2026.

[31] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014. doi: 10.1111/biom.12242.

[32] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, 2017.

[33] Jared Strauch and Amir Asiaee. Improving drug sensitivity prediction and inference by multi-task learning. *bioRxiv*, 2024. doi: 10.1101/2024.05.09.593186.

[34] Jared Strauch, Leila Azinfar, Heather H. Pua, James P. Long, Kevin R. Coombes, and Amir Asiaee. Tissue-aware elastic net decomposition reveals shared and lineage-specific drug response biomarkers. *bioRxiv*, 2026. doi: 10.64898/2026.06.21.733619.

[35] Joost van Rosmalen, David Dejardin, Yvette van Norden, Bob Löwenberg, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27(10):3167–3182, 2018. doi: 10.1177/0962280217694506.

[36] 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, 2018. doi: 10.1080/01621459.2017.1319839.

A Additional derivations

This appendix separates the algebraic borrowing bounds from the representation-risk decomposition. Appendix A.1 proves the scalar and finite-basis information results used in the main text. Appendices A.3 and A.5 give GP-limit and encoder-uncertainty extensions. Appendix A.6 contains the PAC-Bayes and IPM-calibration details for Theorem 2.

A.1 Scalar and finite-basis information proofs

A.1.1 Proof of Corollary 1 and posterior mean

Integrating over b gives $\widehat{\bar{\tau}}^o \mid \bar{\tau} \sim \mathcal{N}(\bar{\tau}, \sigma_o^2/n^o + \sigma_b^2)$. The posterior precision for a normal mean with independent normal observations and prior is the sum of prior and likelihood precisions, hence

$$V_n^{-1} = v_0^{-1} + \frac{n^r}{\sigma_r^2} + \frac{1}{\sigma_o^2/n^o + \sigma_b^2}. \tag{12}$$

Equating the OS precision $(\sigma_o^2/n^o + \sigma_b^2)^{-1}$ with $n^o_{\text{eff}}/\sigma_o^2$ gives $n^o_{\text{eff}} = \sigma_o^2/(\sigma_o^2/n^o + \sigma_b^2) = n^o/(1 + n^o\sigma_b^2/\sigma_o^2)$, and $\lim_{n^o\to\infty} n^o_{\text{eff}} = \sigma_o^2/\sigma_b^2$.

The corresponding posterior mean is

$$m_n = V_n \left\{ \frac{m_0}{v_0} + \frac{n^r \widehat{\bar{\tau}}^r}{\sigma_r^2} + \frac{\widehat{\bar{\tau}}^o}{\sigma_o^2/n^o + \sigma_b^2} \right\}, \tag{13}$$

which makes the borrowing weight explicit: the observational point estimate can be extremely precise for $\bar{\tau} + b$, but its weight for $\bar{\tau}$ is capped by $1/\sigma_b^2$.

A.1.2 Proof of Theorem 1

The marginal covariance of the OS contrast vector is $\Omega = \sigma^2 I + \Phi^o \Sigma_b (\Phi^o)^\top$. The Fisher information for β_τ is $I^o = (\Phi^o)^\top \Omega^{-1} \Phi^o$. If Φ^o has full column rank, the GLS information identity yields

$$(\Phi^o)^\top (\sigma^2 I + \Phi^o \Sigma_b (\Phi^o)^\top)^{-1} \Phi^o = \left\{ \Sigma_b + \sigma^2 ((\Phi^o)^\top \Phi^o)^{-1} \right\}^{-1}. \tag{14}$$

Since $\sigma^2((\Phi^o)^\top\Phi^o)^{-1}$ is positive semidefinite, $\Sigma_b^{-1} - I^o$ is also positive semidefinite. If the empirical feature covariance grows so that $((\Phi^o)^\top\Phi^o)^{-1} \to 0$, then the information converges to Σ_b^{-1}.

A.2 Full-pooling and RCT-only limits

We first study the two extremes of the comparative-bias prior.

Proposition 1 (Full-pooling and RCT-only limits)**.** *Consider the Gaussian B-CALM model in Eqs. (5)–(6) with fixed encoders and proper priors on μ and τ. If $b_\Delta \equiv 0$, the OS and RCT share the same treatment-effect surface and the OS likelihood contributes directly to the posterior precision for τ. If the prior on b_Δ converges to a flat Gaussian process prior under a uniform flat-precision condition on the OS contrast directions, the OS likelihood identifies only $\tau + b_\Delta$ and contributes no finite information to τ beyond information about shared baseline and representation parameters.*

A.3 Gaussian-process limit in Proposition 1

We state the Gaussian-process version in a form that separates two limits. The first is the flat comparative-bias limit, in which the OS contrast is absorbed by an increasingly diffuse prior on b_Δ. The second is the proper GP discretization limit, in which a finite mesh approximation converges to the joint GP model with fixed $\sigma_\Delta^2 < \infty$.

Proof sketch. For fixed z, the OS treated-control contrast is a noisy observation of $\tau(z) + b_\Delta(z)$. When $b_\Delta(z) \equiv 0$, this is a noisy observation of $\tau(z)$. When $b_\Delta(z)$ has an improper flat prior independent of $\tau(z)$, the transformation $(\tau, b_\Delta) \mapsto (\tau, c = \tau + b_\Delta)$ yields an OS likelihood depending on c but not on τ, so integrating over c leaves the marginal posterior for τ proportional to the RCT likelihood times the prior, up to terms involving shared baseline parameters. The GP version below requires a uniform flat-precision condition on the OS contrast directions: if C_m is the covariance induced by the mesh approximation to b_Δ, then $\|(\sigma^2_{\Delta,m} C_m)^{-1}\|_{\text{op}} \to 0$ yields L^1 convergence of the OS marginal-likelihood ratio to one under the RCT-only posterior, hence total-variation convergence. Pointwise variance inflation is not sufficient when the mesh dimension grows, because high-frequency directions can retain nonvanishing precision (Remark 2).

Proposition 2 (Flat GP comparative-bias limit)**.** *Let $\mathcal{Z}_0 \subset \mathbb{R}^d$ be compact. Conditional on fixed encoders and on baseline terms, suppose the Gaussian contrast likelihoods can be written as*

$$Y^r = H^r\tau + \varepsilon^r, \qquad \varepsilon^r \sim \mathcal{N}(0, \Omega^r), \tag{15}$$

$$Y^o = H^o\tau + B_m b_m + \varepsilon^o, \qquad \varepsilon^o \sim \mathcal{N}(0, \Omega^o), \tag{16}$$

where Ω^r and Ω^o are positive definite, H^r and H^o are finite collections of bounded linear evaluation or quadrature functionals on $C(\mathcal{Z}_0)$, and $b_m = (b_\Delta(z_1), \ldots, b_\Delta(z_m))^\top$ is evaluated on a mesh $Z_m = \{z_1, \ldots, z_m\} \subset \mathcal{Z}_0$. Assume $b_m \sim \mathcal{N}(m_m, \sigma^2_{\Delta,m} K_b^{(m)})$, where $K_b^{(m)} = k_b(Z_m, Z_m)$ and k_b is continuous, bounded, and strictly positive definite on $\mathcal{Z}_0$. Let

$$C_m = B_m K_b^{(m)} B_m^\top. \tag{17}$$

Assume that C_m is positive definite on the OS contrast subspace and that the flat-precision condition

$$\left\| (\sigma^2_{\Delta,m} C_m)^{-1} \right\|_{\mathrm{op}} \longrightarrow 0 \tag{18}$$

holds. If the prior Π_τ for τ is a proper Gaussian measure on $C(\mathcal{Z}_0)$, independent of b_m, then the marginal posterior $\Pi_m(\tau \mid Y^r, Y^o)$ converges in total variation to the RCT-only posterior

$$\Pi^r(d\tau \mid Y^r) \propto \varphi_{\Omega^r}\{Y^r - H^r\tau\}\, \Pi_\tau(d\tau). \tag{19}$$

Proof. After integrating b_m, the OS marginal likelihood for τ is

$$L_m^o(\tau) = \varphi_{\Omega^o + \sigma^2_{\Delta,m} C_m}\{Y^o - H^o\tau - B_m m_m\}. \tag{20}$$

Let $V_m = \Omega^o + \sigma^2_{\Delta,m} C_m$ and define the likelihood ratio

$$r_m(\tau) = \frac{L_m^o(\tau)}{L_m^o(0)}. \tag{21}$$

Because Ω^o is positive definite and Eq. (18) holds, $\|V_m^{-1}\|_{\mathrm{op}} \to 0$. For every fixed τ for which $H^o\tau$ is finite,

$$\log r_m(\tau) = (H^o\tau)^\top V_m^{-1}(Y^o - B_m m_m) - \frac{1}{2}(H^o\tau)^\top V_m^{-1}(H^o\tau) \longrightarrow 0. \tag{22}$$

The vector $H^o\tau$ is Gaussian under the proper GP prior Π_τ, since H^o contains finitely many bounded linear functionals. Gaussian exponential moments imply that, for all large m, $r_m(\tau)$ is dominated by an integrable function under the RCT-only posterior $\Pi^r(\cdot \mid Y^r)$. Hence $r_m \to 1$ in $L^1\{\Pi^r(\cdot \mid Y^r)\}$ by dominated convergence.

The marginal posterior with the OS likelihood included has density

$$\frac{d\Pi_m}{d\Pi^r}(\tau) = \frac{r_m(\tau)}{\int r_m(u)\, \Pi^r(du \mid Y^r)}. \tag{23}$$

Since $r_m \to 1$ in $L^1(\Pi^r)$, the denominator converges to one and

$$\|\Pi_m - \Pi^r\|_{\mathrm{TV}} = \frac{1}{2}\int \left| \frac{r_m(\tau)}{\int r_m(u)\, \Pi^r(du \mid Y^r)} - 1 \right| \Pi^r(d\tau \mid Y^r) \longrightarrow 0. \tag{24}$$

The same argument applies after integrating proper Gaussian nuisance parameters, such as μ and b_0, because Fubini's theorem applies to the joint Gaussian prior and the Gaussian likelihood is nonnegative and finite. □

Remark 1 (Kernel and encoder conditions). The preceding proof uses only finite-dimensional covariance matrices, but the following standard conditions are sufficient for the mesh construction to be a discretization of a GP on the latent support. The support $\mathcal{Z}_0$ is compact; the posterior encoder distributions for Z^r and Z^o are supported on $\mathcal{Z}_0$ and have densities bounded above with respect to Lebesgue measure; the mesh fill distance tends to zero; and k_b is continuous, bounded, and admits a Mercer expansion

$$k_b(z, z') = \sum_{j=1}^{\infty} \lambda_j e_j(z) e_j(z'), \tag{25}$$

with $\lambda_j > 0$, $\sum_j \lambda_j < \infty$, and uniformly bounded continuous eigenfunctions. These assumptions make the GP a tight Borel probability measure on $C(\mathcal{Z}_0)$ when the sample paths are continuous, and they give convergence of finite collections of evaluation or stable quadrature functionals. The density bound on the encoder distributions is convenient for replacing empirical mesh sums by integrals over P_Z^r and P_Z^o; it can be weakened to weak convergence of those quadrature rules against bounded continuous functions.

Remark 2 (Why pointwise flatness is not enough). Condition (18) is a uniform flat-precision condition on the OS contrast directions. Pointwise variance inflation does not suffice when the mesh dimension grows. To see this, take a diagonal Mercer representation with eigenvalues $\lambda_j = j^{-2}$ and let the mth discretization retain the first m basis coefficients. Set $\sigma^2_{\Delta,m} = m$. For each fixed coordinate j, $\sigma^2_{\Delta,m}\lambda_j \to \infty$, so the prior is pointwise diffuse on every fixed finite coordinate. Along the new coordinate $j = m$, however, the variance is $\sigma^2_{\Delta,m}\lambda_m = 1/m \to 0$. An OS contrast observation in that coordinate has marginal variance $\sigma_o^2 + 1/m$, not an exploding variance, and therefore supplies nonvanishing information about the corresponding coefficient of τ. Thus the posterior cannot converge uniformly to the RCT-only posterior over the growing cylinders. Requiring $\sigma^2_{\Delta,m}\lambda_{\min}(C_m) \to \infty$, or equivalently Eq. (18), rules out this high-frequency escape.

Proposition 3 (Proper GP mesh limit)**.** *Keep the setup of Proposition 2, but fix $0 < \sigma^2_\Delta < \infty$. Let $b_\Delta \sim \mathcal{GP}(m_\Delta, \sigma^2_\Delta k_b)$ on $C(\mathcal{Z}_0)$, and suppose the mesh approximation satisfies*

$$B_m m_m \to H_b m_\Delta, \qquad B_m K_b^{(m)} B_m^\top \to H_b K_b H_b^\top \tag{26}$$

in Euclidean norm for the finite OS functionals H_b. Then the marginal posterior for τ under the mesh model converges in total variation to the posterior under the joint GP model in which

$$Y^o \mid \tau \sim \mathcal{N}\left\{H^o\tau + H_b m_\Delta,\ \Omega^o + \sigma^2_\Delta H_b K_b H_b^\top\right\}. \tag{27}$$

Consequently, with a proper comparative-bias GP prior, the mesh limit is not the RCT-only posterior. It is the bias-limited posterior that treats the OS contrast as information about $\tau + b_\Delta$ and caps its information about τ through the covariance $\sigma^2_\Delta k_b$.

Proof. Let $L_m^{o,\mathrm{pr}}(\tau)$ be the proper-prior OS marginal likelihood obtained by integrating b_m. By Eq. (26), the mean vector and covariance matrix in $L_m^{o,\mathrm{pr}}(\tau)$ converge to the corresponding mean vector and covariance matrix of the exact GP marginal likelihood $(L^o)^{\mathrm{GP}}(\tau)$ for every fixed τ. The limiting covariance is positive definite because Ω^o is positive definite. The Gaussian density is continuous in its mean and covariance on positive-definite covariance matrices, and the same Gaussian exponential-moment domination used in Proposition 2 yields

$$L_m^{o,\mathrm{pr}}(\tau) \to (L^o)^{\mathrm{GP}}(\tau) \quad \text{in } L^1\{\Pi^r(\cdot \mid Y^r)\}. \tag{28}$$

Writing each posterior density with respect to $\Pi^r(\cdot \mid Y^r)$ and normalizing gives total-variation convergence. The displayed exact-GP likelihood is precisely the likelihood obtained by integrating the proper GP b_Δ, so all OS contrast information enters through the inflated covariance $\Omega^o + \sigma^2_\Delta H_b K_b H_b^\top$. □

Remark 3 (Lifting over encoder uncertainty). The proposition statements bound the conditional posterior given a fixed encoder draw $\phi = (\phi^r, \phi^o)$. Let Π_m^ϕ and $\Pi^{\phi,r}$ denote the corresponding conditional posteriors for τ, and let $q(d\phi)$ be the encoder posterior. If $q(d\phi)$ is independent of the comparative-bias prior structure used in the propositions, and $\|\Pi_m^\phi - \Pi^{\phi,r}\|_{\mathrm{TV}} \to 0$ holds uniformly in ϕ or with a q-integrable dominating function, then Fubini-Tonelli gives convergence of the marginal posterior over (τ, ϕ) to the RCT-only marginal posterior over (τ, ϕ):

$$\left\|\Pi_m^\phi(d\tau)q(d\phi) - \Pi^{\phi,r}(d\tau)q(d\phi)\right\|_{\mathrm{TV}} \le \int \|\Pi_m^\phi - \Pi^{\phi,r}\|_{\mathrm{TV}} q(d\phi) \longrightarrow 0. \tag{29}$$

If uniformity fails but pointwise convergence and a q-integrable bound on $\|\Pi_m^\phi - \Pi^{\phi,r}\|_{\mathrm{TV}}$ hold, dominated convergence gives the same total-variation statement.

One concrete sufficient condition for the dominating bound used in the conditional Gaussian likelihood-ratio proof is that $q(d\phi)$ places mass on encoders with uniformly bounded operator norms $\|H^{\phi,r}\|$ and $\|H^{\phi,o}\|$, uniformly bounded image $\|H^{\phi,o}\tau\|$ over the support of Π_τ, and covariance eigenvalue bounds in the flat or proper mesh argument that are uniform in ϕ.

A.4 Encoder uncertainty in the effective sample size

The preceding results condition on a fixed representation. When the OS encoder is uncertain, posterior variation in the mapped OS contrast acts like an additional irreducible discrepancy term. This is not the usual bias-variance tradeoff; it is extra uncertainty about how the OS contrast maps onto the RCT latent coordinate system.

Proposition 4 (Encoder-aware effective sample size)**.** *Consider the scalar model $\widehat{\bar\tau}^r \mid \bar\tau \sim \mathcal{N}(\bar\tau, \sigma_r^2/n^r)$ and $\widehat{\bar\tau}^o \mid \bar\tau, b, \phi^o \sim \mathcal{N}\{\bar\tau + b + \xi(\phi^o), \sigma_o^2/n^o\}$, where $b \sim \mathcal{N}(0, \sigma_b^2)$ and $\xi(\phi^o) \sim$*

$\mathcal{N}(0, \sigma^2_{\text{enc}})$ *are independent. Then the OS effective sample size becomes*

$$n_{\text{eff}}{}^{o}_{\text{enc}} = \frac{n^o}{1 + n^o(\sigma_b^2 + \sigma^2_{\text{enc}})/\sigma_o^2}, \qquad \lim_{n^o \to \infty} n_{\text{eff}}{}^{o}_{\text{enc}} = \frac{\sigma_o^2}{\sigma_b^2 + \sigma^2_{\text{enc}}}. \tag{30}$$

Encoder uncertainty is therefore additive to comparative-bias uncertainty in the borrowing cap.

A.5 Proof of Proposition 4

Proof. Integrating over the independent Gaussian variables b and $\xi(\phi^o)$ gives

$$\widehat{\bar{\tau}}^o \mid \bar{\tau} \sim \mathcal{N}\left(\bar{\tau},\, \frac{\sigma_o^2}{n^o} + \sigma_b^2 + \sigma^2_{\text{enc}}\right). \tag{31}$$

The RCT contrast, the OS contrast, and the normal prior for $\bar{\tau}$ are conditionally independent normal contributions to the same mean parameter, so posterior precision is prior precision plus likelihood precisions:

$$V_n^{-1} = v_0^{-1} + \frac{n^r}{\sigma_r^2} + \frac{1}{\sigma_o^2/n^o + \sigma_b^2 + \sigma^2_{\text{enc}}}. \tag{32}$$

Equating the OS precision with $n_{\text{eff}}{}^{o}_{\text{enc}}/\sigma_o^2$ gives the displayed effective sample size, and taking $n^o \to \infty$ gives the limit. □

Remark 4 (Estimating the encoder-aware ESS). In a B-CALM fit, $\widehat{\sigma}^2_{\text{enc}}$ can be estimated as the variance of $\widehat{\bar{\tau}}^o$ contributed by posterior draws of ϕ^o, with the observed data, bias draw, and outcome-surface draws held fixed. Plugging this estimate, together with estimates of σ_o^2 and σ_b^2, into Proposition 4 gives a single empirical diagnostic $\widehat{n_{\text{eff}}}{}^{o}_{\text{enc}}$ for borrowing health in the scalar linear-Gaussian setting.

A.6 PAC-Bayes and IPM-calibration details

A.6.1 Definitions and calibration residual

We first make the calibration residual explicit, using the notation of Eqs. (9)–(10). For a full B-CALM draw $f = (\phi^r, \phi^o, \mu, \tau, b_0, b_\Delta)$, let $P^r_Z(f)$ and $P^o_Z(f)$ be the distributions of the latent variables induced by the trial and observational covariate distributions through q_{ϕ^r} and q_{ϕ^o}. Write $\eta^r_f(a,z) = \mu_f(z) + a\tau_f(z)$ and $\eta^o_{\text{deb},f}(a,z) = \eta^o_\star(a,z) - b_{0,f}(z) - ab_{\Delta,f}(z)$. Let $\pi^r(a \mid z)$ be the target trial treatment distribution used to define the risk, for instance the randomized trial distribution or the uniform interventional distribution over $a \in \{0,1\}$.

Let $\eta^r_\star(a,z)$ and $\eta^o_\star(a,z)$ denote the true source-specific conditional outcome-mean surfaces. Write $\mathcal{L}_\ell(t,m) = \mathbb{E}_{Y\sim P_m}\ell(t,Y)$. For a loss $\ell(t,y)$ and two outcome-mean parameters u and v, define the loss discrepancy

$$d_\ell(u,v) = \sup_t \left|\mathcal{L}_\ell(t,u) - \mathcal{L}_\ell(t,v)\right|, \tag{33}$$

where P_m is the conditional outcome distribution indexed by m. In Gaussian squared-error models with clipped predictions or in Bernoulli log-loss models with probabilities bounded away from zero and one, $d_\ell(u,v) \le C_\ell|u-v|$ for a finite constant C_ℓ.

The calibration residual is

$$\varepsilon_{\text{cal}}(\rho) = \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P^r_Z(f)} \sum_{a=0}^{1} \pi^r(a \mid Z)\, d_\ell\left\{\eta^r_\star(a,Z), \eta^o_{\text{deb},f}(a,Z)\right\}. \tag{34}$$

It is nonnegative by construction. It is zero under the noiseless-calibration ideal

$$\eta^r_\star(a,z) = \eta^o_{\text{deb},f}(a,z) \quad \text{for } a \in \{0,1\} \tag{35}$$

for ρ-almost every f and $P^r_Z(f)$-almost every z.

A.6.2 Deterministic calibration and alignment step

Lemma 1 (Calibration and IPM step)**.** *For each f, define the debiased OS conditional risk*

$$\widetilde{r}^o_f(z) = \sum_{a=0}^{1} \pi^r(a \mid z)\, \mathcal{L}_\ell\{\eta^r_f(a,z), \eta^o_{\text{deb},f}(a,z)\}. \tag{36}$$

Assume that $z \mapsto \widetilde{r}^o_f(z)/L$ belongs to the class of test functions $\mathcal{G}$ for ρ-almost every f. Then

$$R^r(\rho) \le \widetilde{R}^o(\rho) + L\Delta_Z(\rho) + \varepsilon_{\text{cal}}(\rho), \tag{37}$$

where

$$\widetilde{R}^o(\rho) = \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P_Z^o(f)}\widetilde{r}_f^o(Z), \tag{38}$$

$$\Delta_Z(\rho) = \mathbb{E}_{f\sim\rho}\,\mathrm{IPM}_{\mathcal{G}}\{P_Z^r(f), P_Z^o(f)\}. \tag{39}$$

Proof. For fixed f and z, the definition of d_ℓ gives

$$\sum_{a=0}^{1}\pi^r(a\mid z)\mathcal{L}_\ell\{\eta_f^r(a,z),\eta_\star^r(a,z)\} \tag{40}$$

$$\le \widetilde{r}_f^o(z) + \sum_{a=0}^{1}\pi^r(a\mid z)d_\ell\left\{\eta_\star^r(a,z),\eta_{\mathrm{deb},f}^o(a,z)\right\}. \tag{41}$$

Integrating this inequality over $Z\sim P_Z^r(f)$ and then over $f\sim\rho$ gives the calibration part. The IPM definition and the assumption $\widetilde{r}_f^o/L\in\mathcal{G}$ give

$$\mathbb{E}_{P_Z^r(f)}\widetilde{r}_f^o(Z) \le \mathbb{E}_{P_Z^o(f)}\widetilde{r}_f^o(Z) + L\ \mathrm{IPM}_{\mathcal{G}}\{P_Z^r(f),P_Z^o(f)\}. \tag{42}$$

Another integration over $f\sim\rho$ proves Eq. (37). Fubini's theorem applies because the loss is bounded in Theorem 2; in the sub-Gaussian variant below it applies after truncation and then monotone convergence. □

Remark 5 (Sufficient conditions for $\widetilde{r}_f^o/L\in\mathcal{G}$). Let $v_{a,f}(z)=\eta_{\mathrm{deb},f}^o(a,z)$, and suppose $\pi^r(a\mid z)$ is constant in z, as in a randomized or uniform target distribution. For squared loss with conditional variance not depending on z, suppose $\|\eta_f^r(a,\cdot)\|_\infty$, $\|\eta_\star^o(a,\cdot)\|_\infty$, $\|b_{0,f}\|_\infty$, and $\|b_{\Delta,f}\|_\infty$ are at most M, and the same four maps are L_η-Lipschitz in z for ρ-almost every f. Then, for $a\in\{0,1\}$,

$$\left|\{\eta_f^r(a,z)-v_{a,f}(z)\}^2-\{\eta_f^r(a,z')-v_{a,f}(z')\}^2\right| \le 2(4M)(4L_\eta)\|z-z'\| = 32ML_\eta\|z-z'\|. \tag{43}$$

Thus $\widetilde{r}_f^o$ is $32ML_\eta$-Lipschitz, since $\sum_a\pi^r(a)=1$. If $\pi^r(a\mid z)$ varies with z, the product-rule terms involving the Lipschitz constants of $\pi^r(a\mid z)$ and the suprema of the arm-specific risks must also be included.

For Bernoulli log-loss, suppose the predicted probabilities $p_{a,f}(z)=\eta_f^r(a,z)$ and the debiased probabilities $v_{a,f}(z)$ lie in $[\epsilon,1-\epsilon]$, where $0<\epsilon<1/2$, and the component maps defining $p_{a,f}$ and $v_{a,f}$ are L_η-Lipschitz. For $h(p,v)=-v\log p-(1-v)\log(1-p)$,

$$\left|\frac{\partial h}{\partial p}\right| = \left|\frac{p-v}{p(1-p)}\right| \le \frac{1}{\epsilon(1-\epsilon)}, \qquad \left|\frac{\partial h}{\partial v}\right| = \left|\log\frac{1-p}{p}\right| \le \log\frac{1-\epsilon}{\epsilon}. \tag{44}$$

Also $\mathrm{Lip}(p_{a,f})\le L_\eta$ and $\mathrm{Lip}(v_{a,f})\le(2+a)L_\eta\le 3L_\eta$, so each arm risk is Lipschitz with constant

$$\frac{L_\eta}{\epsilon(1-\epsilon)} + 3L_\eta\log\frac{1-\epsilon}{\epsilon}. \tag{45}$$

The same constant applies to $\widetilde{r}_f^o$ under a constant target distribution. For a bounded-Lipschitz test-function unit ball, choose L at least the relevant Lipschitz constant, and at least the sup-norm bound if the class also enforces $\|g\|_\infty\le 1$.

Theorem 3 (IPM-calibration target-risk decomposition)**.** *Under the assumptions of Lemma 1, every posterior ρ satisfies*

$$R^r(\rho) \le \widetilde{R}^o(\rho) + L\Delta_Z(\rho) + \varepsilon_{\mathrm{cal}}(\rho). \tag{46}$$

Proof. This is Eq. (37) from Lemma 1. □

A.6.3 Proof of Theorem 2

Proof of Theorem 2. If ρ is not absolutely continuous with respect to ρ_0, then $\mathrm{KL}(\rho\|\rho_0)=\infty$ and the claim is trivial. For absolutely continuous ρ, the McAllester PAC-Bayes inequality for losses in $[0,1]$ states that, with probability at least $1-\delta$ over the RCT sample, all posteriors ρ satisfy

$$\mathrm{kl}\{\widehat{R}^r(\rho),R^r(\rho)\} \le \frac{\mathrm{KL}(\rho\|\rho_0)+\log(2\sqrt{n^r}/\delta)}{n^r}, \tag{47}$$

where $\mathrm{kl}(u,v)$ is the binary relative entropy. The elementary inequality $\mathrm{kl}(u,v)\ge 2(u-v)^2$ for $v\ge u$ gives

$$R^r(\rho) \le \widehat{R}^r(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\rho_0)+\log(2\sqrt{n^r}/\delta)}{2n^r}}. \tag{48}$$

This proves the first term inside the minimum in Eq. (11). The IPM-calibration decomposition is Theorem 3. Since both inequalities hold for the same target risk, the target risk is bounded by the better of the two displays. □

Remark 6 (When the calibration residual vanishes with alignment). Suppose $d_\ell(u,v) \le C_\ell|u-v|$, $\mathcal{G}$ contains the unit ball of bounded Lipschitz functions on $\mathcal{Z}$, and the function

$$g_f(z) = \sum_{a=0}^{1} \pi^r(a \mid z) d_\ell\{\eta^r_\star(a,z), \eta^o_{\mathrm{deb},f}(a,z)\} \tag{49}$$

satisfies $g_f/L_{\mathrm{cal}} \in \mathcal{G}$ for ρ-almost every f. If the ELBO penalty $\widehat{\mathcal{A}}$ is a consistent estimator of $\Delta_Z(\rho)$ and the OS calibration fit gives

$$\mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P^o_Z(f)} g_f(Z) \to 0, \tag{50}$$

then

$$\varepsilon_{\mathrm{cal}}(\rho) \le \mathbb{E}_{f\sim\rho}\mathbb{E}_{Z\sim P^o_Z(f)} g_f(Z) + L_{\mathrm{cal}}\Delta_Z(\rho) \to 0 \tag{51}$$

whenever $\Delta_Z(\rho) \to 0$. This is the precise sense in which alignment alone is not enough: the debiased OS surface must also be calibrated on the shared support.

A.6.4 Sub-Gaussian variant

Proposition 5 (Sub-Gaussian PAC-Bayes variant)**.** *Replace the bounded-loss assumption in Theorem 2 by the following condition: for every f in the support of the data-independent prior ρ_0, the random variable $\ell^r(f;\xi) - R^r(f)$ is σ^2-sub-Gaussian under the RCT distribution, where σ^2 is the tail proxy. Also assume the deterministic IPM-calibration conditions of Lemma 1, with L, $\Delta_Z(\rho)$, $\widetilde{R}^o(\rho)$, and $\varepsilon_{\mathrm{cal}}(\rho)$ defined there. Then, with probability at least $1-\delta$ over the RCT sample, every posterior ρ simultaneously satisfies the sub-Gaussian RCT-only PAC-Bayes bound*

$$R^r(\rho) \le \widehat{R}^r(\rho) + \sigma\sqrt{\frac{2\{\mathrm{KL}(\rho\|\rho_0) + \log(2\sqrt{n^r}/\delta)\}}{n^r}}, \tag{52}$$

where $\mathrm{KL}(\rho\|\rho_0)$ is the Kullback-Leibler divergence from ρ to ρ_0, and the deterministic IPM-calibration decomposition

$$R^r(\rho) \le \widetilde{R}^o(\rho) + L\Delta_Z(\rho) + \varepsilon_{\mathrm{cal}}(\rho). \tag{53}$$

Consequently the target risk is bounded by

$$R^r(\rho) \le \min\left\{ \widehat{R}^r(\rho) + \sigma\sqrt{\frac{2\{\mathrm{KL}(\rho\|\rho_0) + \log(2\sqrt{n^r}/\delta)\}}{n^r}}, \widetilde{R}^o(\rho) + L\Delta_Z(\rho) + \varepsilon_{\mathrm{cal}}(\rho) \right\}. \tag{54}$$

The rate remains $\mathcal{O}((n^r)^{-1/2})$; the cost is the tail constant σ and any constants needed to verify sub-Gaussianity for the chosen loss.

Proof. For fixed f, independence of the RCT sample and the sub-Gaussian assumption imply

$$\mathbb{E}_{\mathcal{D}^r} \exp\left[\lambda\{R^r(f) - \widehat{R}^r(f)\}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2n^r}\right). \tag{55}$$

Integrating with respect to the prior ρ_0 and applying Markov's inequality gives, with probability at least $1-\delta$,

$$\mathbb{E}_{f\sim\rho_0} \exp\left[\lambda\{R^r(f) - \widehat{R}^r(f)\} - \frac{\lambda^2\sigma^2}{2n^r}\right] \le \delta^{-1}. \tag{56}$$

The Donsker-Varadhan variational inequality then gives, simultaneously for all ρ,

$$R^r(\rho) - \widehat{R}^r(\rho) \le \frac{\mathrm{KL}(\rho\|\rho_0) + \log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2n^r}. \tag{57}$$

Optimizing over $\lambda > 0$ yields the displayed square-root term with $\log(1/\delta)$. Replacing this logarithm by the slightly larger McAllester-style $\log(2\sqrt{n^r}/\delta)$ gives Eq. (52). The deterministic IPM-calibration decomposition in Eq. (53) is Theorem 3. Its proof uses Lemma 1, whose argument depends on the loss discrepancy d_ℓ and the condition $\widetilde{r}^o_f/L \in \mathcal{G}$, not on bounded loss; for sub-Gaussian losses the required integrations follow by the truncation and monotone-convergence argument described in the lemma proof. Since both inequalities hold for the same target risk $R^r(\rho)$, the combined bound is Eq. (54). □

A.6.5 Connection to bias-limited information

Remark 7 (Connection to Theorem 1). Consider the local Gaussian linearization in Theorem 1. Let β_τ have prior covariance Σ_τ, let the RCT contribution to the local posterior precision be I^r, and let the OS contribution be I^o. Theorem 1 gives that both I^o and $\Sigma_b^{-1} - I^o$ are positive semidefinite. If $\Lambda^r = \Sigma_\tau^{-1} + I^r$ and $\Lambda^{r,o} = \Lambda^r + I^o$, then the covariance part of the Gaussian complexity satisfies

$$\frac{1}{2}\left[\log\det(\Lambda^{r,o}) - \log\det(\Lambda^r)\right] \tag{58}$$

$$= \frac{1}{2}\log\det(I + (\Lambda^r)^{-1/2} I^o (\Lambda^r)^{-1/2}) \tag{59}$$

$$\leq \frac{1}{2}\log\det(I + (\Lambda^r)^{-1/2} \Sigma_b^{-1} (\Lambda^r)^{-1/2}). \tag{60}$$

Thus, apart from the usual posterior-mean term, which is controlled by the same prior-norm clipping or localization assumptions used in PAC-Bayes with unbounded parameter spaces, the OS can increase the PAC-Bayes complexity by at most the displayed constant. If $\Sigma_b = \sigma_\Delta^2 I_d$, this constant is bounded by

$$\frac{d}{2}\log\left\{1 + \frac{\lambda_{\max}((\Lambda^r)^{-1})}{\sigma_\Delta^2}\right\}. \tag{61}$$

The additional contribution to the first square-root term in Eq. (11) is therefore at most this constant divided by $2n^r$ inside the square root, up to the stated mean-localization term. This is the PAC-Bayes counterpart of the bias-limited information bound: adding the OS likelihood cannot create more treatment-effect precision than the comparative-bias prior permits.

Sub-Gaussian losses. The bounded-loss assumption can be replaced by a uniform sub-Gaussian tail condition: for every draw f, the centered RCT loss $U_f = \ell\{\eta_f^r(A^r, Z^r), Y^r\} - \mathbb{E}\ell\{\eta_f^r(A^r, Z^r), Y^r\}$ satisfies $\mathbb{E}\{\exp(\lambda U_f)\} \leq \exp(\lambda^2\sigma^2/2)$ for all real λ. This is a tail condition on the loss, not an assumption that the outcome itself is Gaussian; ordinary unbounded squared loss requires clipping, localization, or stronger moment conditions to satisfy it. Appendix A.6 gives the same $\mathcal{O}((n^r)^{-1/2})$ rate with a different constant under this condition.

A.7 Remarks on the assumptions

Alignment can be encouraged in several ways. If the shared covariate block W is informative, learned encoders can be constrained to preserve it, for example by adding an auxiliary decoder that predicts W from Z. If linked records are available, such as trial participants linked to external registry or EHR records, one can directly penalize disagreement between the two encoded views of the same unit. In a learned-encoder implementation, the IPM of Eq. (3) could also be used as a regularizer.

Here "source-specific" means that the two latent distributions are generated from different observed covariate systems and different encoders: trial units pass through q_{ϕ^r} and OS units pass through q_{ϕ^o}. Assumption 3 is therefore the mathematical version of "the two covariate systems can be put on a common latent coordinate system, but not perfectly." The remaining mismatch is not ignored; it appears as the IPM term Δ_Z. Small alignment error makes OS information more useful, while large alignment error pushes the analysis toward RCT-only information.

A.8 Binary-outcome notation

For binary outcomes, let g be a generalized-linear-model link function and let $h = g^{-1}$. The same bias decomposition can be placed on the linear-predictor scale

$$\eta^r(a, z) = \mu(z) + a\tau(z), \tag{62}$$

$$\eta^o(a, z) = \mu(z) + a\tau(z) + b_0(z) + ab_\Delta(z), \tag{63}$$

with

$$\mathbb{P}(Y^s = 1 \mid A = a, Z = z) = h\{\eta^s(a, z)\}. \tag{64}$$

For a logit link, $h = \text{expit}$. In that case $\tau(z)$ is a conditional log-odds contrast, not a risk-difference CATE. Posterior risks for trial units are

$$\rho_a(x^r) = \mathbb{E}\left[h\{\mu(Z) + a\tau(Z)\} \mid Z \sim q_{\phi^r}(\cdot \mid x^r), \mathcal{D}\right], \tag{65}$$

and the reported risk-difference CATE is

$$\tau_{RD}(x^r) = \rho_1(x^r) - \rho_0(x^r). \tag{66}$$

Prior calibration for b_Δ should be interpreted in the units induced by the chosen link function; for example, under a logit link it is a prior on the OS-versus-RCT log-odds-ratio bias. We do not evaluate binary outcomes in this paper.

A.9 Reporting recommendations for the bias priors

Because b_Δ, and with one-arm external controls also b_0, is not identified without prior restrictions, its prior scale is a sensitivity parameter. We recommend four reporting rules.

(1) Report the whole curve. Report posterior summaries over a prespecified grid of scales running from full pooling to trial-only inference, rather than at one selected scale. A model-averaged or score-selected scale can be shown as one point on that curve, but because the bias is not identified it is not a borrowing strength learned from the data.

(2) Elicit the scale in outcome units. For the continuous-outcome setting emphasized in the main text, a pointwise normal prior statement such as

$$\mathbb{P}\{|b_\Delta(z)| < \delta\} = 0.95 \tag{67}$$

corresponds to $\sigma_\Delta \approx \delta/1.96$. If the analyst believes the OS treatment contrast may be biased by as much as 0.4 outcome units in either direction, then $\sigma_\Delta \approx 0.20$; equivalently, $\sigma_\Delta = 0.2$ treats a 0.2-unit gap between the trial and OS contrasts as a one-standard-deviation surprise. The grid should be prespecified, for example $\delta \in \{0, 0.2, 0.4, 0.8, \infty\}$, and a robust mixture prior allows posterior retreat from the optimistic component.

(3) Report the tipping point. Report the smallest scale on the grid at which the substantive conclusion changes, for example the first scale at which the 90% interval for the ATE includes zero. A finite grid establishes the conclusion only at the evaluated scales. When the posterior is Gaussian, as in the real-data application, compute interval limits in closed form: the Monte Carlo error of a posterior quantile can exceed the margin by which an interval excludes zero.

(4) Report a distributional-discrepancy diagnostic. Next to the posterior, report the maximum mean discrepancy (MMD) [14] between the trial and OS samples on the representation the outcome model uses. With features standardized by the pooled mean and standard deviation, a Gaussian kernel $k(u, v) = \exp\{-\|u - v\|^2/(2h^2)\}$, and bandwidth h equal to the median pairwise distance in the pooled sample, the unbiased estimate is

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^r(n^r-1)} \sum_{i \neq i'} k(z_i^r, z_{i'}^r) + \frac{1}{n^o(n^o-1)} \sum_{j \neq j'} k(z_j^o, z_{j'}^o) - \frac{2}{n^r n^o} \sum_{i,j} k(z_i^r, z_j^o), \tag{68}$$

with a permutation p-value as a secondary summary. The statistic measures distributional discrepancy; it is not evidence of overlap or of adequate confounding adjustment. It equals the IPM of Theorem 2 only when the class of test functions $\mathcal{G}$ is the unit ball of the kernel's reproducing-kernel Hilbert space, the theorem additionally requires its regularity condition, and when computed on shared covariates rather than on a latent representation it is not an estimate of Δ_Z. Appendix D.6 reports it for the synthetic alignment ladder and the real-data application.

A.10 Data governance and broader impacts

The real-data analysis uses previously collected pediatric-obesity trial and EHR control data under the source studies' data-use and institutional-review constraints; no new participants were recruited and no crowdworkers were used. Any participant-risk, consent, and disclosure procedures were handled under those source-study reviews. Patient-level data are not released. The intended benefit is more cautious external-control augmentation for small trials, especially when observational cohorts are large but imperfectly comparable. The main risk is overinterpretation: calibrated borrowing can quantify sensitivity to bias priors, but it does not remove the need for an explicitly trial-defined estimand, domain review, and study-specific validation before clinical or policy use.

A.11 Future directions

Proposition 4 gives an encoder-aware extension of Corollary 1 in the linear-Gaussian case, adding the encoder-offset variance to the comparative-bias variance in the OS effective sample size. Several extensions follow directly. (i) *Time-varying treatments*: extending the bias decomposition to longitudinal settings where the comparative bias may differ across treatment regimes raises identifiability questions tied to the structural-nested model literature. (ii) *Decision-aware borrowing*: replacing the symmetric prior on b_Δ with a decision-theoretic prior that penalizes errors in the direction that would change a treatment recommendation. (iii) *Multi-source fusion*: more than one observational study with study-specific bias functions and a hierarchical prior across sources.

A.12 Reproducibility

Design grids and seeds. The factorial main simulation uses trial sizes $n^r \in \{250, 500\}$, observational size $n^o = 5{,}000$, and 50 seeds per cell over the 24-cell condition grid of Appendix B.2. The saturation experiment uses $n^r \in \{250, 500\}$, $n^o \in \{500, 1{,}000, 2{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000\}$, fixed $\sigma_\Delta \in \{0.1, 0.2, 0.4, \infty\}$, and 50 seeds per cell. The semi-synthetic benchmark uses $n^r = 500$, $n^o = 5{,}000$, and 50 seeds.

Prior-scale settings across studies. The main factorial grid selects σ_Δ per replicate by empirical-Bayes model averaging (Algorithm 2; prior mean $m_\Delta = 0$). The semi-synthetic and calibration studies use the same empirical-Bayes averaging. The saturation experiment fixes σ_Δ per curve and sets the comparative-bias prior mean to $m_\Delta = 0.2$, which matches the location of the true bias template; its coverage values are conservative and should not be read as evidence about the adaptive procedure. The real-data application holds σ_Δ diffuse and selects σ_0 as described in Appendix B.6.

Baseline implementations. CF (the Wager causal forest) uses double machine learning with random-forest nuisance models. The Hahn Bayesian causal forest (BCF) is run only in the real-data analysis; in our implementation it uses two BART components (prognostic f and moderating τ) with explicit PGBART step methods, 50 trees per component, and 500 posterior draws after 250 tuning steps; it omits the propensity-score input of Hahn et al. [16], and predictions at new points use the nearest training point. The R-OSCAR, MR-OSCAR, and CALM baselines reuse the parent calibrated-borrowing implementations with a 200-replicate trial bootstrap for intervals: the reported estimate is the pointwise bootstrap median, with a symmetric interval whose width is the bootstrap 5–95% range. Our implementation splits covariates by position, treating all but the last two columns as shared. In the synthetic designs this makes x_1, x_2, x_3 shared and x_4, x_5 source-specific (x_3 has the same loading in both sources except in the mismatch-heavy regime); in the real-data analysis, where every covariate is shared, it treats the last two encoded columns as source-specific. The three methods' grid means agree to within 10^{-7}; Appendix D.4 reports them at full precision.

Compute and code. All results were produced on a single workstation. Code for all experiments, with the synthetic and semi-synthetic result files, simulation seeds, hyperparameter configurations, an ordered reproduction notebook, and figure-generation scripts, is available at `https://github.com/AsiaeeLab/b-calm`. The real-data analysis code is included, but per-source IRB and data-use constraints prevent release of its input files and of patient-level data.

B Experimental details

This appendix gives the details needed to reproduce the experiments in Section 6: metric definitions (Appendix B.1), the synthetic data-generating process and condition grid (Appendix B.2), the B-CALM estimator exactly as fit in all experiments (Appendix B.4), the semi-synthetic construction routine (Appendix B.5), and the real-data cohort and configuration (Appendix B.6). Software, seeds, and compute are collected in Appendix A.12.

B.1 Metrics

Each simulated replicate uses an evaluation sample $x_1, \ldots, x_m$ with $m = 600$ fresh draws from the trial covariate distribution ($m = 500$ in the calibration study and $m = 400$ in the saturation study) and true CATEs $\tau(x_j)$ known by construction. For a method reporting posterior-mean CATE $\widehat{\tau}(\cdot)$ and posterior standard deviation $\widehat{s}(\cdot)$, the pointwise 90% interval is $\widehat{\tau}(x_j) \pm z\,\widehat{s}(x_j)$ with $z = z_{0.95} \approx 1.645$, and we record

$$\mathrm{RMSE} = \left\{ \tfrac{1}{m} \textstyle\sum_j \{\widehat{\tau}(x_j) - \tau(x_j)\}^2 \right\}^{1/2}, \qquad \mathrm{Cov} = \tfrac{1}{m} \textstyle\sum_j \mathbb{1}\{|\widehat{\tau}(x_j) - \tau(x_j)| \le z\,\widehat{s}(x_j)\}, \tag{69}$$

and $\mathrm{Width} = \frac{1}{m} \sum_j 2z\,\widehat{s}(x_j)$. Negative transfer in a replicate is the indicator $\mathbb{1}\{\mathrm{RMSE}_{\mathrm{method}} > \mathrm{RMSE}_{\text{RCT-only}}\}$ computed against the RCT-only fit of the same replicate; the reported NT is its average over replicates. For B-CALM the posterior of $\tau(x)$ is Gaussian with closed-form moments (Appendix B.4), so $\widehat{\tau}$ and $\widehat{s}$ are analytic; for the empirical-Bayes mixture, we use the exact mixture mean and variance. CF reports asymptotic influence-function standard errors; the OSCAR/CALM family reports intervals from a 200-replicate trial bootstrap.

B.2 Synthetic data-generating process

Latent state and covariates. For each unit, draw a latent state $z = (z_1, z_2, z_3)^\top \sim \mathcal{N}(0, I_3)$. Observational units receive a regime-specific mean shift $z \mapsto z + \delta$ with $\delta = (0.05, -0.02, 0.00)$

(shared-only), $\delta = (0.25, -0.15, 0.10)$ (balanced), or $\delta = (0.60, -0.40, 0.25)$ (mismatch-heavy). Five observed covariates are generated as noisy linear maps of z with independent noise $\epsilon_k \sim \mathcal{N}(0, s^2)$, where $s = 0.10, 0.25, 0.45$ by regime. The first two covariates are common to both sources and form the shared block W:

$$x_1 = z_1 + \epsilon_1, \qquad x_2 = z_2 + \epsilon_2. \tag{70}$$

The remaining three covariates have source-specific loadings κ^s:

$$x_3 = \kappa_3^s\, z_3 + \epsilon_3, \qquad x_4 = \kappa_4^s\, z_1 + \epsilon_4, \qquad x_5 = \kappa_5^s\, z_2 + \epsilon_5, \tag{71}$$

with, per regime,

	κ_3		κ_4		κ_5	
Regime	RCT	OS	RCT	OS	RCT	OS
shared-only*	1.0	1.0	$0.6z_1 - 0.2z_2$		0.4	0.4
balanced	1.0	1.0	0.70	0.45	0.25	0.40
mismatch-heavy	0.90	0.35	0.90	0.15	0.10	0.80

*In the shared-only regime both sources use the same maps ($x_4 = 0.6z_1 - 0.2z_2 + \epsilon_4$ and $x_5 = 0.4z_3 + \epsilon_5$), so all five covariates are effectively shared; the balanced and mismatch-heavy regimes make x_3, x_4, x_5 progressively more source-specific, which creates the trial-only block U and OS-only block V.

Outcome surfaces. Outcome complexity is tied to the regime: shared-only uses linear surfaces, balanced adds mild nonlinearity, and mismatch-heavy adds further nonlinearity. With covariates $x = (x_1, \ldots, x_5)$,

$$\mu(x) = 0.25 + 0.45x_1 - 0.25x_2 + 0.18x_3 + 0.08x_4 + \underbrace{0.12 \sin x_1 - 0.08x_2x_3}_{\text{balanced, mismatch-heavy}} + \underbrace{0.10 \cos x_4 + 0.08 \tanh(x_1x_5)}_{\text{mismatch-heavy}}, \tag{72}$$

$$\tau(x) = 0.35 + 0.16x_1 - 0.10x_3 + \underbrace{0.08 \sin x_2 + 0.05x_1x_4}_{\text{balanced, mismatch-heavy}} + \underbrace{0.10 \tanh x_5 - 0.06 \cos(x_1 + x_3)}_{\text{mismatch-heavy}}. \tag{73}$$

Bias functions and treatment assignment. Under the no-bias regime, $b_0 = b_\Delta \equiv 0$. Otherwise the OS has baseline bias $b_0(x) = 0.20 + 0.15x_1 - 0.10x_3$, and under the comparative regime additionally

$$b_\Delta(x) = \frac{s_\Delta}{0.2} \left\{0.20 + 0.12x_1 - 0.08x_2 + 0.04 \tanh x_4\right\}, \tag{74}$$

where s_Δ is the comparative-bias strength ($s_\Delta = 0.2$ except in the sweep, which uses $s_\Delta \in \{0, 0.2, 0.5, 1.0\}$); the baseline-only regime sets $b_\Delta \equiv 0$. Trial treatment is assigned by balanced randomization ($\lfloor n^r/2 \rfloor$ treated, randomly permuted). OS treatment is confounded:

$$A^o \sim \text{Bernoulli}\left[\text{expit}\{-0.05 + 0.85x_1 - 0.55x_2 + 0.35x_4 + 0.35x_3\, \mathbb{1}(\text{comparative})\}\right]. \tag{75}$$

Outcomes are $Y = \mu(X) + A\,\tau(X) + \mathbb{1}(\text{OS})\{b_0(X) + A\,b_\Delta(X)\} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$.

Condition grid and evaluation. The factorial grid crosses the three regimes with the three bias settings at $s_\Delta = 0.2$, plus the balanced-regime comparative sweep $s_\Delta \in \{0, 0.5, 1.0\}$, each at $n^r \in \{250, 500\}$ with $n^o = 5{,}000$: 24 cells, 50 seeds per cell. Every replicate is evaluated on $m = 600$ fresh trial-distribution covariate draws with known $\tau(\cdot)$. OS-only is omitted from the three sweep-extension cells.

B.3 Learned-encoder implementation

Variational implementation. For computation, we approximate the posterior over functions by basis-function or neural-network parameterizations, with weight vectors $\omega_\mu, \omega_\tau, \omega_0, \omega_\Delta$ for μ, τ, b_0, b_Δ respectively. Let $\theta = (\omega_\mu, \omega_\tau, \omega_0, \omega_\Delta, \phi^r, \phi^o)$ collect these weights together with the encoder parameters. Given priors $p(\theta)$, a variational approximation $q_\lambda(\theta)$ maximizes

$$\mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda}\left[\sum_{i=1}^{n^r} \log p_{\text{G}}\{Y_i^r; \eta^r(A_i^r, Z_i^r), \sigma_r^2\} + \sum_{j=1}^{n^o} \log p_{\text{G}}\{Y_j^o; \eta^o(A_j^o, Z_j^o), \sigma_o^2\}\right] - \text{KL}\{q_\lambda(\theta) \| p(\theta)\} - \lambda_{\text{align}}\, \widehat{\mathcal{A}}(q_{\phi^r}, q_{\phi^o}), \tag{76}$$

Algorithm 1 Synthetic data generation (one replicate)

Require: regime $r \in \{\text{shared-only, balanced, mismatch-heavy}\}$, bias setting, strength s_Δ, sizes n^r, n^o, seed

1: draw latent states $z \sim \mathcal{N}(0, I_3)$ for n^r trial and n^o OS units; shift OS states by δ_r
2: map to covariates $x_1, \ldots, x_5$ with regime- and source-specific loadings and noise scale s_r
3: assign trial treatment by balanced randomization; draw OS treatment from Eq. (75)
4: compute $\mu(x), \tau(x)$ at regime complexity; compute $b_0(x), b_\Delta(x)$ per bias setting and Eq. (74)
5: draw $Y = \mu + A\tau + \mathbb{1}(\text{OS})(b_0 + Ab_\Delta) + \mathcal{N}(0, 1)$
6: draw $m = 600$ evaluation covariates from the trial law; store true $\tau(\cdot)$ there

where $Z_i^r \sim q_{\phi^r}(\cdot \mid W_i^r, U_i^r)$ and $Z_j^o \sim q_{\phi^o}(\cdot \mid W_j^o, V_j^o)$. For learned-encoder implementations, the alignment penalty $\widehat{\mathcal{A}}$ may be an MMD penalty, reconstruction loss for shared covariates, adversarial source-classification loss, or a combination; in the fixed-basis experiments, this term is absent because the representation is chosen before fitting. We compute the posterior by Monte Carlo draws of (θ, Z) and transformation through Eq. (7).

Algorithm sketch. B-CALM follows five steps: (1) choose an encoder family and, when encoders are learned, an alignment loss; (2) choose priors for μ, τ, b_0, b_Δ, including prespecified sensitivity values for σ_Δ; (3) optimize the evidence lower bound (ELBO) in Eq. (76); (4) draw posterior samples of the trial control surface, treated surface, and CATEs for trial-population units; and (5) report posterior summaries and sensitivity curves over σ_Δ.

B.4 B-CALM implementation and pseudocode

Model. All experiments use the closed-form Gaussian instance of Section 4 with identity encoders: each source's covariate vector is its latent state, and b_0 and b_Δ absorb the remaining source discrepancy. Fix a basis map $\phi(\cdot) \in \mathbb{R}^k$ and write $\mu(x) = \phi(x)^\top \omega_\mu$, $\tau(x) = \phi(x)^\top \omega_\tau$, $b_0(x) = \phi(x)^\top \omega_0$, $b_\Delta(x) = \phi(x)^\top \omega_\Delta$. Stacking $\theta = (\omega_\mu, \omega_\tau, \omega_0, \omega_\Delta) \in \mathbb{R}^{4k}$, each trial row contributes the design vector $(\phi^\top, a\phi^\top, 0, 0)$ and each OS row $(\phi^\top, a\phi^\top, \phi^\top, a\phi^\top)$, with Gaussian likelihood $Y \mid \cdot \sim \mathcal{N}(\text{design} \cdot \theta, \sigma_y^2)$, $\sigma_y = 1$. The basis matches the regime: linear $\phi(x) = (1, x)$ for shared-only; adding $\sin x_1$, x_2^2, $x_1 x_3$ for balanced; adding $\cos x_4$, $\tanh(x_1 x_5)$ for mismatch-heavy. The basis overlaps but does not span the true surfaces (for example, the balanced-regime truth contains $x_2 x_3$, $\sin x_2$, and $x_1 x_4$, none of which is in the basis), so the outcome model is mildly misspecified by design.

Priors and posterior. Coefficients have independent Gaussian priors: $\omega_\mu, \omega_\tau \sim \mathcal{N}(0, 3^2 I_k)$, $\omega_0 \sim \mathcal{N}(0, \sigma_0^2 I_k)$ with $\sigma_0 = 0.6$, and $\omega_\Delta \sim \mathcal{N}(m_\Delta e_1, \sigma_\Delta^2 I_k)$, where e_1 selects the intercept coefficient; $m_\Delta = 0$ in the factorial grid and semi-synthetic study and $m_\Delta = 0.2$ in the saturation experiment. Diffuse and degenerate scales are implemented numerically (prior variance 10^{10} for ∞, 10^{-10} for 0). With design matrix Φ and prior precision P, the posterior is Gaussian with precision $\Lambda = P + \Phi^\top \Phi / \sigma_y^2$ and mean $\Lambda^{-1}(P m_{\text{prior}} + \Phi^\top Y / \sigma_y^2)$, computed by Cholesky factorization. The posterior CATE at x is Gaussian with mean $\phi(x)^\top \widehat{\omega}_\tau$ and variance $\phi(x)^\top \text{Cov}(\omega_\tau) \phi(x)$; no sampling is required.

Empirical-Bayes selection of σ_Δ. In the factorial grid, σ_Δ is selected per replicate by model averaging over the fixed candidate set $S = \{0.01, 0.05, 0.2, 0.5, 1.0, \infty\}$ with fixed prior weights $\pi = (0.01, 0.01, 0.42, 0.21, 0.13, 0.22)$. The weights are asymmetric by design: an overly tight σ_Δ over-borrows a biased contrast and threatens validity, while an overly diffuse σ_Δ only costs efficiency, so the prior stays skeptical on the tight scales, keeps a mild mode at moderate borrowing, and carries a heavy diffuse tail for retreat — the robust-mixture structure recommended in Appendix A.9. The weights are one global choice, used unchanged in every cell and study. They were fixed during method development after examining performance on the simulation designs, so they are not prespecified in the confirmatory sense; Appendix D.10 reports the main grid under two alternative weight vectors. Per-replicate posterior weights are exported with the results for inspection. Each candidate is scored by an auxiliary Gaussian model for fitted contrast summaries, not by the marginal likelihood of the full joint model: fit the RCT-only model (diffuse bias priors) and an OS-only contrast model, extract the two posterior contrast-coefficient distributions $(\widehat{\tau}^r, \Sigma^r)$ and $(\widehat{\tau}^o, \Sigma^o)$, model their difference as $\widehat{\tau}^o - \widehat{\tau}^r \sim \mathcal{N}\{0, \Sigma^r + \Sigma^o + (v_\sigma + 10^{-9})I\}$, and score $\log p_\sigma = \log \mathcal{N}\{\widehat{\tau}^o - \widehat{\tau}^r;\ 0,\ \Sigma^r + \Sigma^o + (v_\sigma + 10^{-9})I\}$, where $v_\sigma = \max(\sigma^2, 10^{-8})$ for finite σ and $v_\infty = 10^4$. The diffuse candidate is scored with this large finite variance (the joint fit itself uses prior variance 10^{10}), so that it keeps positive weight

Algorithm 2 B-CALM, Gaussian finite-basis instance with empirical-Bayes averaging

Require: trial data $\mathcal{D}^r$, OS data $\mathcal{D}^o$, basis ϕ, candidate set S, prior weights π, $\sigma_0 = 0.6$, $\sigma_y = 1$

1: build design rows $(\phi, a\phi, 0, 0)$ for trial units and $(\phi, a\phi, \phi, a\phi)$ for OS units
2: fit RCT-only ($\sigma_\Delta = \sigma_0 = \infty$, trial rows only) and OS-only contrast ($\sigma_\Delta = 0$, OS rows only); extract contrast posteriors $(\widehat{\tau}^r, \Sigma^r)$, $(\widehat{\tau}^o, \Sigma^o)$
3: **for** $\sigma \in S$ **do**
4: fit the joint Gaussian posterior with comparative-bias scale σ (closed form, Cholesky)
5: compute $\log p_\sigma = \log \mathcal{N}\{\widehat{\tau}^o - \widehat{\tau}^r; 0, \Sigma^r + \Sigma^o + (v_\sigma + 10^{-9})I\}$, with $v_\sigma = \max(\sigma^2, 10^{-8})$ and $v_\infty = 10^4$
6: **end for**
7: set weights $w_\sigma \propto \pi_\sigma \exp(\log p_\sigma)$; return the weighted mixture of fits
8: report posterior CATE mean/variance at any x from exact mixture moments; intervals are mean ± 1.645 sd

and the average can retreat to trial-only borrowing when the contrasts conflict; the term $10^{-9}I$ is a numerical jitter, as is a $10^{-8}I$ term added to the posterior precision before factorization. Posterior model weights are $w_\sigma \propto \pi_\sigma \exp(\log p_\sigma)$, and the reported posterior is the corresponding mixture of Gaussian fits with exact mixture moments. The semi-synthetic and calibration studies use the same averaging; the saturation experiment instead fixes $\sigma_\Delta \in \{0.1, 0.2, 0.4, \infty\}$ per curve, and the semi-synthetic sensitivity figure varies fixed σ_Δ over $\{0, 0.1, 0.2, 0.4, \infty\}$.

Baseline configurations of the same model. RCT-only fits the trial rows alone with diffuse bias priors; OS-only fits the OS rows alone with $\sigma_0 = \sigma_\Delta = 0$ (its contrast is used directly); pooled fits both sources with $\sigma_0 = \sigma_\Delta = 0$, forcing $b_0 = b_\Delta \equiv 0$.

B.5 Semi-synthetic construction modeled on STAR

The benchmark keeps STAR-type covariates and marginals but imposes known outcome surfaces, following Asiaee et al. [5]. For source $s \in \{r, o\}$, covariates are drawn as sex $\sim$ Bernoulli(0.49); race (Black) $\sim$ Bernoulli(0.32) for the trial and Bernoulli(0.38) for the OS; free lunch $\sim$ Bernoulli(0.42) vs. Bernoulli(0.50); small class $\sim$ Bernoulli(0.33) vs. Bernoulli(0.24); teacher experience $\sim \mathcal{N}(8, 3^2)$ vs. $\mathcal{N}(7, 3^2)$; baseline score $= \mathcal{N}(0, 1) - 0.35\,\text{freelunch} + 0.018\,\text{teacherexp}$; and school quality $= \mathcal{N}(0, 0.7^2) + 0.2(1 - \text{freelunch})$. The seven covariates enter centered and scaled as $x = (\text{baseline score},\ \text{sex} - 0.5,\ \text{race} - 0.32,\ \text{freelunch} - 0.45,\ \text{smallclass} - 0.30,\ (\text{teacherexp} - 8)/3,\ \text{school quality})$. The imposed surfaces are

$$\mu(x) = 0.15 + 0.60x_1 + 0.12x_2 - 0.22x_4 + 0.16x_6 + 0.20x_7 + 0.08 \sin x_1, \tag{77}$$

$$\tau(x) = 0.22 + 0.18(1 - x_4) + 0.12x_5 + 0.08 \tanh x_1, \tag{78}$$

with OS bias functions $b_0(x) = 0.22 + 0.18x_4 - 0.12x_7$ and $b_\Delta(x) = 0.18 + 0.10x_4 - 0.08x_5 + 0.05 \tanh x_1$ (comparative strength 0.2). Trial treatment is balanced-randomized; OS treatment is confounded via $A^o \sim \text{Bernoulli}[\text{expit}\{-0.15 + 0.75x_1 - 0.55x_4 + 0.30x_7\}]$; outcomes add $\mathcal{N}(0, 1)$ noise. Sizes are $n^r = 500$, $n^o = 5{,}000$, with 50 seeds and 600 trial-law evaluation draws. B-CALM uses the star basis ($\phi = (1, x, \sin x_1, x_2^2, x_1x_3, \cos x_4, \tanh(x_1x_5))$) with the same empirical-Bayes averaging over σ_Δ as the main grid; the sensitivity analysis in Figure 4, which varies fixed σ_Δ from 0 to ∞, shows how conclusions move across the whole borrowing range. Among the calibrated-borrowing family, MR-OSCAR is run in this study and represents the family. The sensitivity figure fixes one seed and traces posterior CATEs at seven representative covariate profiles: the evaluation units at seven evenly spaced positions in the ordering by baseline score.

B.6 Real-data cohort and configuration

The trial is a multi-site pediatric-obesity prevention RCT with $n^r = 385$ children and randomized treatment; the external cohort is a cross-sectional EHR sample of $n^o = 15{,}150$ children drawn from the same health systems, none of whom received the trial intervention, so every EHR row enters with $A = 0$ (an external-control convention). The outcome is the 12-month weight-for-length z-score. Six shared covariates are available in both sources: site, sex, race/ethnicity, language, insurance, and baseline weight-for-length z-score; five trial-only covariates (household food security, parent education, income, health literacy, and birth weight) are retained for sensitivity analyses only. B-CALM is fit on the six shared covariates with a linear basis. With one-arm external controls the comparative-bias scale σ_Δ is unidentified and held diffuse; borrowing is governed by the baseline-commensurability scale σ_0, selected over the candidate set $\{0, 0.025, 0.05, 0.1, 0.2, 0.5, \infty\}$ with

fixed prior weights $(0.05, 0.20, 0.30, 0.20, 0.10, 0.05, 0.10)$ by the prior-weighted score $\log \pi(\sigma_0) - \log\{\text{interval width}(\sigma_0)\} - 0.02\,|\widehat{\text{ATE}}(\sigma_0)|$, which is maximized at $\sigma_0 = 0.05$. The skeptical scale $\sigma_0 = 0.6$ used in the bias-susceptibility diagnostic is the default baseline-bias prior of the synthetic experiments, an order of magnitude above the selected value. BCF is excluded from the bias-susceptibility diagnostic because that diagnostic requires a grid of refits (roughly $2(K+1)$ for K strata), which is infeasible at BCF's MCMC cost; it appears in the main real-data comparison, Table 15. Patient-level data cannot be released (Appendix A.10).

C Full main simulation results table

The 138-row main-results table summarized by Table 1 in the body is reproduced in full below. Each row reports the average across 50 seeds for one method in one combination of (DGP, bias regime, comparative-bias strength, n^r), using the condition grid of Appendix B.2; RMSE is reported as mean $\pm$ across-seed standard deviation. Bold and underline preserve the generated per-cell highlight markers for quick scanning.

Table 3: Full main simulation results over 50 seeds per cell. Within each cell we bold the method with the lowest negative-transfer rate and underline the method with coverage closest to the 90% target from below. Pooled and CF (the Wager causal forest) achieve lower RMSE under no bias because pooling is variance-optimal there, but their coverage collapses under comparative bias. R-OSCAR, MR-OSCAR, and CALM differ by less than 10^{-4} in every cell and are shown as one OSCAR/CALM row (full precision in Appendix D.4); the family sits just below nominal coverage on average while declining to borrow under bias. The Hahn Bayesian causal forest (BCF) is run only in the real-data analysis. Bias-limited width saturation is guaranteed by Corollary 1.

DGP	Bias	b_Δ	n^r	Method	Coverage	Neg. transfer	RMSE	Width
balanced	baseline-only		250	pooled	0.794	0.000	0.135 ± 0.031	0.285
balanced	baseline-only		250	OS-only	0.794	0.020	0.137 ± 0.034	0.297
balanced	baseline-only		250	CF	0.967	0.000	0.152 ± 0.016	0.760
balanced	baseline-only		250	**B-CALM**	**0.978**	**0.000**	**0.201 ± 0.076**	**0.795**
balanced	baseline-only		250	OSCAR/CALM	0.918	0.020	0.258 ± 0.075	0.895
balanced	baseline-only		250	RCT-only	0.926	0.000	0.370 ± 0.101	1.225
balanced	baseline-only		500	pooled	0.802	0.020	0.124 ± 0.024	0.274
balanced	baseline-only		500	OS-only	0.807	0.020	0.129 ± 0.030	0.297
balanced	baseline-only		500	B-CALM	0.954	0.000	0.164 ± 0.051	0.591
balanced	baseline-only		500	OSCAR/CALM	0.867	0.000	0.201 ± 0.056	0.608
balanced	baseline-only		500	**CF**	**0.962**	**0.000**	**0.153 ± 0.016**	**0.741**
balanced	baseline-only		500	RCT-only	0.877	0.000	0.276 ± 0.056	0.841
balanced	comparative	0.0	250	pooled	0.765	0.000	0.138 ± 0.031	0.286
balanced	comparative	0.0	250	CF	0.965	0.000	0.155 ± 0.016	0.764
balanced	comparative	0.0	250	**B-CALM**	**0.977**	**0.000**	**0.203 ± 0.076**	**0.795**
balanced	comparative	0.0	250	OSCAR/CALM	0.918	0.020	0.258 ± 0.075	0.895
balanced	comparative	0.0	250	RCT-only	0.926	0.000	0.370 ± 0.101	1.225
balanced	comparative	0.0	500	pooled	0.765	0.020	0.131 ± 0.030	0.275
balanced	comparative	0.0	500	B-CALM	0.952	0.020	0.167 ± 0.049	0.595
balanced	comparative	0.0	500	OSCAR/CALM	0.867	0.000	0.201 ± 0.056	0.608
balanced	comparative	0.0	500	CF	0.963	0.040	0.153 ± 0.018	0.748
balanced	comparative	0.0	500	**RCT-only**	**0.877**	**0.000**	**0.276 ± 0.056**	**0.841**
balanced	comparative	0.2	250	B-CALM	0.924	0.020	0.241 ± 0.060	0.857
balanced	comparative	0.2	250	pooled	0.308	0.180	0.298 ± 0.032	0.287
balanced	comparative	0.2	250	OS-only	0.320	0.220	0.295 ± 0.032	0.299
balanced	comparative	0.2	250	CF	0.791	0.200	0.284 ± 0.033	0.771
balanced	comparative	0.2	250	OSCAR/CALM	0.903	0.000	0.268 ± 0.092	0.903
balanced	comparative	0.2	250	**RCT-only**	**0.911**	**0.000**	**0.382 ± 0.083**	**1.233**
balanced	comparative	0.2	500	OSCAR/CALM	0.903	0.020	0.183 ± 0.047	0.610
balanced	comparative	0.2	500	B-CALM	0.928	0.040	0.192 ± 0.043	0.664
balanced	comparative	0.2	500	pooled	0.309	0.600	0.289 ± 0.032	0.276
balanced	comparative	0.2	500	OS-only	0.330	0.600	0.291 ± 0.036	0.300
balanced	comparative	0.2	500	**RCT-only**	**0.909**	**0.000**	**0.269 ± 0.061**	**0.847**
balanced	comparative	0.2	500	CF	0.796	0.580	0.277 ± 0.031	0.760
balanced	comparative	0.5	250	OSCAR/CALM	0.904	0.000	0.270 ± 0.068	0.910
balanced	comparative	0.5	250	B-CALM	0.859	0.140	0.295 ± 0.083	0.933
balanced	comparative	0.5	250	**RCT-only**	**0.912**	**0.000**	**0.386 ± 0.111**	**1.245**
balanced	comparative	0.5	250	pooled	0.136	0.960	0.663 ± 0.033	0.288
balanced	comparative	0.5	250	CF	0.370	0.960	0.619 ± 0.036	0.798
balanced	comparative	0.5	500	**OSCAR/CALM**	**0.880**	**0.000**	**0.199 ± 0.056**	**0.609**
balanced	comparative	0.5	500	B-CALM	0.857	0.060	0.239 ± 0.055	0.697
balanced	comparative	0.5	500	RCT-only	0.880	0.000	0.293 ± 0.065	0.845
balanced	comparative	0.5	500	CF	0.393	1.000	0.590 ± 0.040	0.786
balanced	comparative	0.5	500	pooled	0.132	1.000	0.636 ± 0.035	0.276
balanced	comparative	1.0	250	OSCAR/CALM	0.870	0.000	0.302 ± 0.087	0.917
balanced	comparative	1.0	250	B-CALM	0.893	0.060	0.346 ± 0.089	1.094
balanced	comparative	1.0	250	**RCT-only**	**0.875**	**0.000**	**0.418 ± 0.095**	**1.245**
balanced	comparative	1.0	250	pooled	0.074	1.000	1.283 ± 0.056	0.287
balanced	comparative	1.0	250	CF	0.224	1.000	1.185 ± 0.057	0.874
balanced	comparative	1.0	500	OSCAR/CALM	0.870	0.020	0.200 ± 0.059	0.611
balanced	comparative	1.0	500	B-CALM	0.897	0.080	0.251 ± 0.072	0.782
balanced	comparative	1.0	500	**RCT-only**	**0.889**	**0.000**	**0.280 ± 0.075**	**0.845**
balanced	comparative	1.0	500	CF	0.229	1.000	1.142 ± 0.044	0.853

continued on next page

(continued)

DGP	Bias	b_Δ	n^r	Method	Coverage	Neg. transfer	RMSE	Width
balanced	comparative	1.0	500	pooled	0.071	1.000	1.226 ± 0.044	0.275
balanced	none		250	pooled	0.799	0.000	0.131 ± 0.029	0.285
balanced	none		250	OS-only	0.794	0.020	0.137 ± 0.034	0.297
balanced	none		250	CF	0.964	0.000	0.154 ± 0.017	0.750
balanced	none		250	**B-CALM**	**0.978**	**0.000**	**0.201 ± 0.076**	**0.795**
balanced	none		250	OSCAR/CALM	0.918	0.020	0.258 ± 0.075	0.895
balanced	none		250	RCT-only	0.926	0.000	0.370 ± 0.101	1.225
balanced	none		500	pooled	0.815	0.000	0.118 ± 0.022	0.274
balanced	none		500	OS-only	0.807	0.020	0.129 ± 0.030	0.297
balanced	none		500	B-CALM	0.954	0.000	0.164 ± 0.051	0.591
balanced	none		500	OSCAR/CALM	0.867	0.000	0.201 ± 0.056	0.608
balanced	none		500	**CF**	**0.965**	**0.000**	**0.151 ± 0.014**	**0.738**
balanced	none		500	RCT-only	0.877	0.000	0.276 ± 0.056	0.841

Avg coverage: B-CALM 0.929; RCT-only 0.899; OSCAR/CALM 0.890; pooled 0.481; OS-only 0.642; CF 0.716
Avg neg-transfer: B-CALM 0.035; RCT-only 0.000; OSCAR/CALM 0.008; pooled 0.398; OS-only 0.150; CF 0.398

mismatch-heavy	baseline-only		250	pooled	0.787	0.000	0.196 ± 0.052	0.434
mismatch-heavy	baseline-only		250	OS-only	0.805	0.000	0.214 ± 0.059	0.507
mismatch-heavy	baseline-only		250	CF	0.952	0.000	0.165 ± 0.020	0.747
mismatch-heavy	baseline-only		250	**B-CALM**	**0.966**	**0.000**	**0.244 ± 0.081**	**0.915**
mismatch-heavy	baseline-only		250	OSCAR/CALM	0.905	0.000	0.280 ± 0.074	0.922
mismatch-heavy	baseline-only		250	RCT-only	0.907	0.000	0.440 ± 0.110	1.374
mismatch-heavy	baseline-only		500	pooled	0.810	0.020	0.165 ± 0.035	0.392
mismatch-heavy	baseline-only		500	OS-only	0.816	0.160	0.217 ± 0.075	0.505
mismatch-heavy	baseline-only		500	B-CALM	0.942	0.000	0.192 ± 0.053	0.667
mismatch-heavy	baseline-only		500	OSCAR/CALM	0.846	0.000	0.221 ± 0.051	0.628
mismatch-heavy	baseline-only		500	**CF**	**0.951**	**0.000**	**0.164 ± 0.018**	**0.740**
mismatch-heavy	baseline-only		500	RCT-only	0.883	0.000	0.307 ± 0.065	0.935
mismatch-heavy	comparative	0.2	250	CF	0.818	0.060	0.262 ± 0.033	0.760
mismatch-heavy	comparative	0.2	250	OSCAR/CALM	0.890	0.040	0.288 ± 0.094	0.928
mismatch-heavy	comparative	0.2	250	B-CALM	0.926	0.020	0.275 ± 0.069	0.976
mismatch-heavy	comparative	0.2	250	pooled	0.404	0.180	0.338 ± 0.045	0.435
mismatch-heavy	comparative	0.2	250	OS-only	0.464	0.220	0.346 ± 0.055	0.507
mismatch-heavy	comparative	0.2	250	**RCT-only**	**0.896**	**0.000**	**0.443 ± 0.106**	**1.385**
mismatch-heavy	comparative	0.2	500	OSCAR/CALM	0.880	0.000	0.205 ± 0.042	0.631
mismatch-heavy	comparative	0.2	500	B-CALM	0.932	0.020	0.216 ± 0.040	0.759
mismatch-heavy	comparative	0.2	500	CF	0.818	0.300	0.258 ± 0.026	0.745
mismatch-heavy	comparative	0.2	500	pooled	0.383	0.600	0.318 ± 0.039	0.395
mismatch-heavy	comparative	0.2	500	**RCT-only**	**0.900**	**0.000**	**0.305 ± 0.063**	**0.940**
mismatch-heavy	comparative	0.2	500	OS-only	0.466	0.660	0.342 ± 0.053	0.510
mismatch-heavy	none		250	pooled	0.797	0.000	0.191 ± 0.057	0.434
mismatch-heavy	none		250	OS-only	0.805	0.000	0.214 ± 0.059	0.507
mismatch-heavy	none		250	CF	0.943	0.000	0.172 ± 0.021	0.739
mismatch-heavy	none		250	**B-CALM**	**0.966**	**0.000**	**0.244 ± 0.081**	**0.915**
mismatch-heavy	none		250	OSCAR/CALM	0.906	0.000	0.280 ± 0.074	0.922
mismatch-heavy	none		250	RCT-only	0.907	0.000	0.440 ± 0.110	1.374
mismatch-heavy	none		500	pooled	0.832	0.020	0.156 ± 0.031	0.392
mismatch-heavy	none		500	OS-only	0.816	0.160	0.217 ± 0.076	0.505
mismatch-heavy	none		500	B-CALM	0.942	0.000	0.192 ± 0.053	0.667
mismatch-heavy	none		500	OSCAR/CALM	0.846	0.000	0.221 ± 0.051	0.628
mismatch-heavy	none		500	**CF**	**0.944**	**0.000**	**0.169 ± 0.022**	**0.731**
mismatch-heavy	none		500	RCT-only	0.883	0.000	0.307 ± 0.065	0.935

Avg coverage: B-CALM 0.946; RCT-only 0.896; OSCAR/CALM 0.879; pooled 0.669; OS-only 0.695; CF 0.904
Avg neg-transfer: B-CALM 0.007; RCT-only 0.000; OSCAR/CALM 0.007; pooled 0.137; OS-only 0.200; CF 0.060

shared-only	baseline-only		250	pooled	0.883	0.000	0.069 ± 0.022	0.224
shared-only	baseline-only		250	OS-only	0.892	0.000	0.071 ± 0.021	0.231
shared-only	baseline-only		250	**B-CALM**	**0.980**	**0.000**	**0.156 ± 0.056**	**0.683**
shared-only	baseline-only		250	CF	0.966	0.080	0.156 ± 0.016	0.759
shared-only	baseline-only		250	OSCAR/CALM	0.918	0.040	0.226 ± 0.076	0.804
shared-only	baseline-only		250	RCT-only	0.936	0.000	0.265 ± 0.078	0.977
shared-only	baseline-only		500	pooled	0.904	0.000	0.064 ± 0.023	0.218
shared-only	baseline-only		500	OS-only	0.915	0.000	0.067 ± 0.021	0.231
shared-only	baseline-only		500	**B-CALM**	**0.944**	**0.000**	**0.137 ± 0.055**	**0.516**
shared-only	baseline-only		500	OSCAR/CALM	0.869	0.000	0.177 ± 0.059	0.548
shared-only	baseline-only		500	CF	0.966	0.240	0.152 ± 0.017	0.748
shared-only	baseline-only		500	RCT-only	0.887	0.000	0.212 ± 0.062	0.692
shared-only	comparative	0.2	250	B-CALM	0.926	0.040	0.187 ± 0.070	0.714
shared-only	comparative	0.2	250	pooled	0.273	0.500	0.266 ± 0.033	0.225
shared-only	comparative	0.2	250	OS-only	0.270	0.500	0.265 ± 0.033	0.232
shared-only	comparative	0.2	250	OSCAR/CALM	0.910	0.060	0.234 ± 0.085	0.813
shared-only	comparative	0.2	250	CF	0.794	0.560	0.280 ± 0.034	0.774
shared-only	comparative	0.2	250	**RCT-only**	**0.916**	**0.000**	**0.276 ± 0.089**	**0.980**
shared-only	comparative	0.2	500	B-CALM	0.930	0.060	0.143 ± 0.045	0.546
shared-only	comparative	0.2	500	OSCAR/CALM	0.921	0.020	0.150 ± 0.051	0.550
shared-only	comparative	0.2	500	**RCT-only**	**0.921**	**0.000**	**0.196 ± 0.050**	**0.694**
shared-only	comparative	0.2	500	pooled	0.277	0.920	0.262 ± 0.028	0.219
shared-only	comparative	0.2	500	OS-only	0.283	0.860	0.260 ± 0.028	0.232
shared-only	comparative	0.2	500	CF	0.791	0.900	0.281 ± 0.034	0.769
shared-only	none		250	pooled	0.895	0.000	0.068 ± 0.020	0.224
shared-only	none		250	OS-only	0.892	0.000	0.071 ± 0.021	0.231
shared-only	none		250	**B-CALM**	**0.980**	**0.000**	**0.156 ± 0.056**	**0.683**
shared-only	none		250	CF	0.962	0.080	0.157 ± 0.017	0.754
shared-only	none		250	OSCAR/CALM	0.918	0.040	0.226 ± 0.076	0.804
shared-only	none		250	RCT-only	0.936	0.000	0.265 ± 0.078	0.977
shared-only	none		500	pooled	0.927	0.000	0.061 ± 0.019	0.218

continued on next page

(continued)

DGP	Bias	b_Δ	n^r	Method	Coverage	Neg. transfer	RMSE	Width
shared-only	none		500	OS-only	0.915	0.000	0.067 ± 0.021	0.231
shared-only	none		500	**B-CALM**	**0.945**	**0.000**	$\mathbf{0.137 \pm 0.054}$	**0.516**
shared-only	none		500	OSCAR/CALM	0.869	0.000	0.177 ± 0.059	0.548
shared-only	none		500	CF	0.965	0.180	0.151 ± 0.015	0.738
shared-only	none		500	RCT-only	0.887	0.000	0.212 ± 0.062	0.692

Avg coverage: B-CALM 0.951; RCT-only 0.914; OSCAR/CALM 0.901; pooled 0.693; OS-only 0.694; CF 0.907
Avg neg-transfer: B-CALM 0.017; RCT-only 0.000; OSCAR/CALM 0.027; pooled 0.237; OS-only 0.227; CF 0.340

D Additional experimental results

D.1 Saturation effective sample size

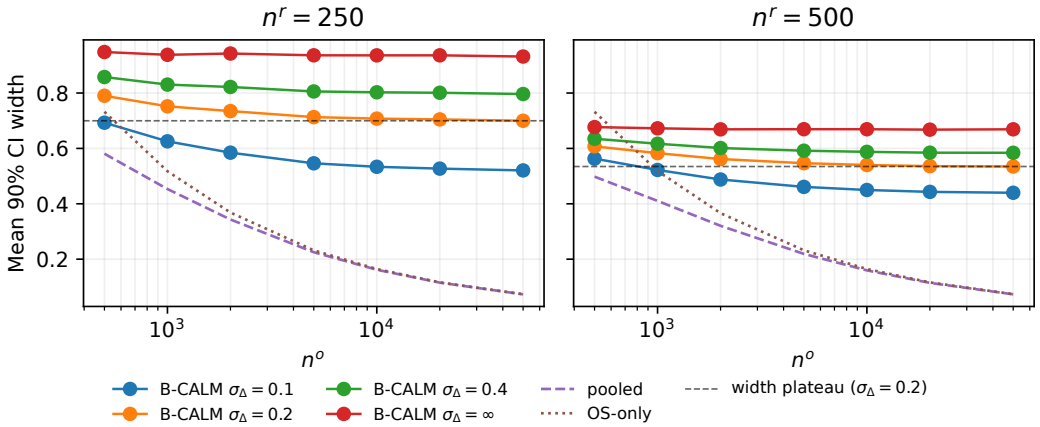


Figure 3: Interval-width saturation: mean 90% credible-interval width as n^o grows with n^r fixed (log horizontal scale; 50 replicates per point). B-CALM interval widths flatten to a σ_Δ-dependent floor (the horizontal dashed reference marks the $\sigma_\Delta = 0.2$ plateau), while the pooled and OS-only curves, which nearly overlap, contract toward zero regardless of bias. Effective-sample-size values appear in Appendix D.1, Table 4.

Table 4: Saturation diagnostic at the largest n^o. Empirical n^o_{eff} is computed from posterior precision after subtracting the RCT contribution.

n^r	σ_Δ	Emp. n^o_{eff}	Th. n^o_{eff}	Width	Cov.
250	0.1	402.7	396.8	0.521	0.989
250	0.2	105.6	99.8	0.700	0.953
250	0.4	30.1	25.0	0.796	0.929
250	inf	4.0	0.0	0.932	0.923
500	0.1	402.0	396.8	0.440	0.975
500	0.2	105.1	99.8	0.535	0.941
500	0.4	29.8	25.0	0.584	0.930
500	inf	3.5	0.0	0.669	0.921

D.2 Semi-synthetic benchmark

Table 5 reports RMSE and 90% coverage over 50 replicates of the STAR-like benchmark; Appendix B.5 specifies the construction routine, covariate marginals, imposed outcome surfaces, and bias functions. Figure 4 gives the accompanying sensitivity analysis: posterior CATEs at seven representative covariate profiles as the fixed comparative-bias scale σ_Δ varies from 0 (pooled borrowing) to ∞ (RCT-only inference).

Table 5: STAR-like semi-synthetic benchmark with known potential outcomes (50 replicates; construction in Appendix B.5). MR-OSCAR represents the OSCAR/CALM family, whose members agree to within 10^{-7} in the factorial study (Appendix D.4); it is the family member run in this study.

Method	CATE RMSE	90% coverage
B-CALM	0.191 ± 0.057	0.936
pooled	0.194 ± 0.025	0.425
OS-only	0.216 ± 0.028	0.377
CF	0.218 ± 0.024	0.856
MR-OSCAR	0.228 ± 0.067	0.889
RCT-only	0.313 ± 0.078	0.888

Table 6: B-CALM ablations on the balanced comparative-bias setting (strength 0.2, $n^r = 500$, 50 seeds). Coverage and negative-transfer metrics are shown before RMSE to emphasize calibrated borrowing.

Variant	Coverage	Neg. transfer	RMSE	Width
Full B-CALM	0.915	0.020	0.195 ± 0.047	0.670
No b_0 ($\sigma_0 = 0$)	0.791	0.280	0.237 ± 0.072	0.614
No b_Δ ($\sigma_\Delta = 0$)	0.324	0.500	0.259 ± 0.026	0.279

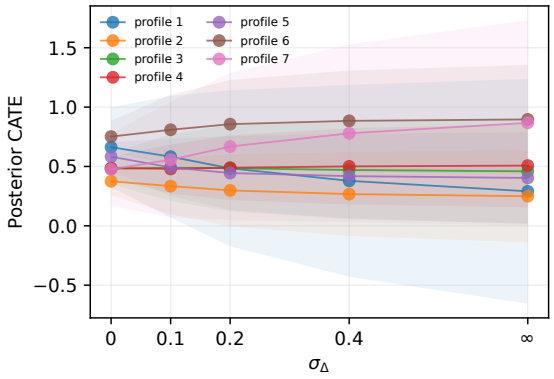


Figure 4: Sensitivity analysis on the STAR-like semi-synthetic benchmark (one replicate). Each line is the posterior CATE at one of seven representative covariate profiles (evaluation units at evenly spaced positions in the ordering by baseline score), with shaded 90% credible bands, as the fixed σ_Δ varies; estimates move between pooled borrowing ($\sigma_\Delta \to 0$) and RCT-only inference ($\sigma_\Delta \to \infty$).

D.3 Ablations

The ablation uses the balanced DGP at comparative-bias strength 0.2, $n^r = 500$, $n^o = 5{,}000$, and 50 seeds per variant. *Full B-CALM* is the empirical-Bayes–averaged fit of Algorithm 2. *No* b_0 forces the baseline shift to zero ($\sigma_0 = 0$) and keeps the same empirical-Bayes averaging over σ_Δ, so the two fits differ only in b_0; *No* b_Δ pools the treatment contrast directly ($\sigma_\Delta = 0$).

D.4 Full-precision results for the OSCAR/CALM family

Table 7 reports R-OSCAR, MR-OSCAR, and CALM separately. All methods share seeds, so any difference between them comes from the code path. The three differ on every seed-cell in RMSE and width, but by at most 6.3×10^{-6}, three orders of magnitude below the Monte Carlo standard error of a grid mean, and their 90% intervals cover the same evaluation points on all but two seed-cells. The reason is structural: the three differ only in how the OS base surface is built (Appendix A.12), and the same trial-anchored calibration, doubly robust pseudo-outcome, and CATE regression follow, so in these designs the calibration absorbs nearly all of the difference. The family is not degenerate: MR-OSCAR differs from trial-only by 0.090 in mean absolute RMSE, with a nonzero difference on every seed-cell, and from pooled by 0.205.

D.5 Results by bias regime

Table 8 regroups the factorial study by bias regime, which the grid average hides. B-CALM's coverage is not uniform across regimes: 0.961 without comparative bias, 0.928 at strength 0.2, 0.858

Table 7: Full-precision grid means of the OSCAR/CALM family over the factorial study (24 cells $\times$ 50 seeds), the Monte Carlo standard error (MC SE) of a grid mean, the largest difference between two members on a single seed-cell, and the number of the 1,200 seed-cells on which the members differ. The last row gives the real-data ATE.

Metric	R-OSCAR	MR-OSCAR	CALM	MC SE	Max gap	Cells
RMSE	0.228486399	0.228486369	0.228486370	0.0020	6.3×10^{-6}	1,200
Coverage	0.890112500	0.890112500	0.890112500	0.0033	1.7×10^{-3}	2
Width	0.741605770	0.741605798	0.741605800	0.0013	3.9×10^{-6}	1,200
NT	0.012500000	0.012500000	0.012500000	0.0032	0	0
Real-data ATE	-0.1784943	-0.1784945	-0.1784215	–	–	–

Table 8: Synthetic factorial study grouped by bias regime: mean 90% coverage, interval width, RMSE, and negative transfer (NT). The no-bias and baseline-only groups pool six cells (three regimes, two trial sizes), the strength-0.2 group pools six, and the strength-0.5 and 1.0 groups pool the two balanced cells of the sweep. B-CALM's coverage standard error is in parentheses. OSCAR/CALM averages its three members.

Method	Coverage	Width	RMSE	NT
No bias				
B-CALM	0.961 (0.004)	0.695	0.182	0.000
OSCAR/CALM	0.887	0.734	0.227	0.010
pooled	0.844	0.305	0.121	0.003
CF	0.957	0.742	0.159	0.043
RCT-only	0.903	1.007	0.312	0.000
Baseline only				
B-CALM	0.961 (0.004)	0.694	0.182	0.000
OSCAR/CALM	0.887	0.734	0.227	0.010
pooled	0.830	0.305	0.126	0.007
CF	0.961	0.749	0.157	0.053
RCT-only	0.903	1.007	0.312	0.000
Comparative, $s_\Delta = 0.2$				
B-CALM	0.928 (0.005)	0.753	0.209	0.033
OSCAR/CALM	0.901	0.739	0.221	0.023
pooled	0.326	0.306	0.295	0.497
CF	0.801	0.763	0.274	0.433
RCT-only	0.909	1.013	0.312	0.000
Comparative, $s_\Delta = 0.5$				
B-CALM	0.858 (0.011)	0.815	0.267	0.100
OSCAR/CALM	0.892	0.759	0.234	0.000
pooled	0.134	0.282	0.649	0.980
CF	0.381	0.792	0.605	0.980
RCT-only	0.896	1.045	0.339	0.000
Comparative, $s_\Delta = 1.0$				
B-CALM	0.895 (0.010)	0.938	0.298	0.070
OSCAR/CALM	0.870	0.764	0.251	0.010
pooled	0.073	0.281	1.255	1.000
CF	0.227	0.864	1.163	1.000
RCT-only	0.882	1.045	0.349	0.000

at 0.5 (3.8 standard errors below 0.90), and 0.895 at 1.0, so it is not monotone in the bias strength. It is never more than 0.042 below nominal, and it over-covers by up to 0.061 without comparative bias. At strength 0.5 the OSCAR/CALM family, which declines to borrow contrast information at every bias level, is better calibrated (0.892) and has lower RMSE; at strength 1.0 B-CALM's coverage is the closer to nominal. B-CALM's negative transfer stays at or below 0.10 in every regime, while pooled and CF reach 0.98–1.00 at strengths 0.5 and 1.0.

D.6 Alignment ladder and distributional discrepancy

The three covariate regimes progressively decouple the source-specific covariate loadings (Appendix B.2), but they also add outcome nonlinearity, covariate noise, and a larger latent shift, so they cannot isolate the effect of alignment. Table 9 reports B-CALM along this ladder together with the MMD diagnostic of Appendix A.9; the permutation test uses all trial units and $2{,}000$ OS units. The

Table 9: Alignment ladder. Top: B-CALM's 90% coverage, width, width relative to trial-only, RMSE, and negative transfer, with pooled coverage for comparison ($n^r = 500$, $n^o = 5{,}000$, 50 seeds). Bottom: maximum mean discrepancy between the trial and OS samples on the five-coordinate working representation and on the covariates x_1, x_2, x_3 that our OSCAR implementation treats as shared; the real-data row uses the six shared covariates as encoded for the model ($n^r = 385$, $n^o = 15{,}150$).

Regime	Coverage	Width	Ratio	RMSE	NT	Pooled cov.
No bias						
Shared-only	0.962	0.600	0.72	0.146	0.000	0.911
Balanced	0.966	0.693	0.67	0.182	0.000	0.807
Mismatch-heavy	0.954	0.791	0.69	0.218	0.000	0.814
Comparative, $s_\Delta = 0.2$						
Shared-only	0.928	0.630	0.75	0.165	0.050	0.275
Balanced	0.926	0.760	0.73	0.216	0.030	0.309
Mismatch-heavy	0.929	0.868	0.75	0.246	0.020	0.394

Trial–OS discrepancy (MMD), mean (SD) over 50 seeds; fraction of seeds with permutation $p < 0.05$

	Working representation		Shared covariates	
Shared-only	0.016 (0.015)	0.18	0.015 (0.015)	0.14
Balanced	0.108 (0.014)	1.00	0.088 (0.014)	1.00
Mismatch-heavy	0.256 (0.010)	1.00	0.234 (0.011)	1.00
Real data	–	–	0.353	$p = 0.005$

discrepancy grows along the ladder, and B-CALM widens with it, by 38% under comparative bias and 32% without, while its coverage and negative transfer stay flat and its width stays a quarter to a third below trial-only. In the real data the trial and EHR samples also differ clearly on the shared covariates (MMD 0.353, permutation $p = 0.005$); the feature spaces differ from the synthetic ones, so the values are not directly comparable.

D.7 Scalar and robust-MAP comparative bias

Hierarchical priors on CALM's shared outcome surfaces do not by themselves separate prognostic from contrast borrowing, produce the information ceiling of Theorem 1, or give a covariate-indexed discrepancy; a hierarchical model obtains these only by placing its prior on the contrast discrepancy, which is b_Δ. To isolate the value of a covariate-indexed discrepancy, we compare B-CALM with the same Gaussian outcome model in which b_Δ is held constant over the covariates, which is what a commensurate or single-scale hierarchical prior on the contrast amounts to: all coefficients of b_Δ except the intercept are pinned to zero (prior variance 10^{-10}), and the intercept receives the same empirical-Bayes average over σ_Δ as B-CALM. A second variant also makes b_0 constant. The robust-MAP variant is a two-component mixture of the scalar model, with intercept prior standard deviations 0.05 and 10 and prior weights $1/2$ each, combined by the joint marginal likelihood; it does not use the six-scale average. The data come from the balanced design with comparative bias that either varies with the covariates (Eq. (74)) or is constant, $b_\Delta \equiv s_\Delta$. With covariate-varying bias, the scalar models borrow confidently through a discrepancy they cannot represent and are worse on every metric; at strength 0.4 they cover 0.40–0.47 against 0.87, with negative transfer 0.84–0.86 against 0.08. With constant bias, the scalar model is the more efficient estimator (RMSE 0.156 against 0.215 at strength 0.4), though it under-covers (0.780 against 0.865).

D.8 QR-learner benchmark

We ran the authors' implementation of the mixed QR-learner [26], unmodified, on the comparisons of Table 10 at strength 0.4, with gradient-boosted outcome models, a cross-validated logistic study classifier, a linear final CATE model, two-fold cross-fitting, and the trial propensity fixed at its randomized value 0.5. The implementation returns point estimates only, so the comparison covers accuracy and negative transfer, not intervals. Negative transfer uses the same trial-only Gaussian fit as Table 10, not the QR-learner's own trial-only learner. B-CALM has lower RMSE and lower negative transfer under both bias shapes (Table 11).

D.9 Go/No-Go operating characteristics

The rule declares Go when the posterior probability that the trial-population ATE is positive is at least 0.975, so its nominal false-Go rate is 0.025. For each covariate regime we compute the population mean of τ under the trial covariate law from a $200{,}000$-unit draw (0.350, 0.386, and 0.376 for shared-only, balanced, and mismatch-heavy) and subtract it from the treatment effect in both sources,

Table 10: Function-valued versus scalar comparative-bias models on the balanced design ($n^r = 500$, $n^o = 5{,}000$, 50 seeds). NT is measured against a trial-only fit of the same Gaussian model on the same replicate.

Method	Coverage	Width	RMSE	NT
Covariate-varying comparative bias, $s_\Delta = 0.2$				
B-CALM (function-valued b_Δ)	0.928	0.664	0.192	0.040
Commensurate (scalar b_Δ)	0.626	0.381	0.218	0.220
Commensurate (scalar b_Δ and b_0)	0.619	0.384	0.225	0.260
Robust MAP (scalar b_Δ)	0.510	0.352	0.235	0.340
Covariate-varying comparative bias, $s_\Delta = 0.4$				
B-CALM (function-valued b_Δ)	0.872	0.685	0.220	0.080
Commensurate (scalar b_Δ)	0.404	0.382	0.347	0.840
Commensurate (scalar b_Δ and b_0)	0.398	0.380	0.356	0.860
Robust MAP (scalar b_Δ)	0.470	0.458	0.362	0.860
Constant comparative bias, $s_\Delta = 0.2$				
B-CALM (function-valued b_Δ)	0.917	0.625	0.178	0.040
Commensurate (scalar b_Δ)	0.788	0.379	0.145	0.040
Commensurate (scalar b_Δ and b_0)	0.781	0.381	0.149	0.040
Robust MAP (scalar b_Δ)	0.625	0.351	0.163	0.040
Constant comparative bias, $s_\Delta = 0.4$				
B-CALM (function-valued b_Δ)	0.865	0.680	0.215	0.060
Commensurate (scalar b_Δ)	0.780	0.395	0.156	0.060
Commensurate (scalar b_Δ and b_0)	0.765	0.394	0.161	0.060
Robust MAP (scalar b_Δ)	0.745	0.460	0.180	0.220

Table 11: QR-learner [26] and B-CALM at comparative-bias strength 0.4 on the balanced design (50 seeds). Both NT columns use the same trial-only reference fit on each replicate.

Comparative bias	QR-learner		B-CALM	
	RMSE	NT	RMSE	NT
Covariate-varying	0.256	0.400	0.220	0.080
Constant	0.250	0.300	0.215	0.060

so the population ATE is 0 under the null; for the alternative we subtract the same mean less 0.2. This keeps all effect heterogeneity and the bias structure, and the realized mean truth on the evaluation sets is within 0.001 of its target. Each cell has $n^r = 500$, OS size 5,000 or 20,000, and 200 replicates. Table 12 gives the full grid for the empirical-Bayes average. Table 13 reports each fixed σ_Δ in that average and shows where the excess false-Go rate under comparative bias comes from: the tight scales $\sigma_\Delta \leq 0.05$ make a false Go in 45–100% of replicates, $\sigma_\Delta = 0.2$ in 3.5–11%, and $\sigma_\Delta \geq 0.5$ keeps the rate at or below 0.035, with power close to that of the trial-only rule. Under intermediate comparative bias the average still places weight on $\sigma_\Delta = 0.2$ (Section 6), which produces its excess.

D.10 Sensitivity to the prior weights on σ_Δ

The weights π on the candidate scales were fixed during method development after examining performance on the simulation designs, and then held unchanged for every reported experiment. Table 14 refits the main grid under uniform weights and under a vector that places more weight on the moderate scale $\sigma_\Delta = 0.2$. The second vector gives nearly the same results. Uniform weights put a third of the prior mass on the two tight scales, which over-borrows under comparative bias: coverage at strength 0.2 falls from 0.928 to 0.784 and the lowest cell coverage from 0.857 to 0.731, while RMSE and width improve where there is no comparative bias. Calibration under comparative bias therefore depends on keeping little prior mass on the tight scales, which is why the weights are skeptical there.

Table 12: Go/No-Go operating characteristics: rate of a Go decision (posterior probability of a positive trial-population ATE at least 0.975) under a null population ATE of 0 (Type I) and an alternative of 0.2 (Power), by covariate regime, bias setting, and OS size $n^o \in \{5{,}000,\ 20{,}000\}$; $n^r = 500$, 200 replicates per cell, so binomial standard errors are at most 0.035 (about 0.011 at the nominal 0.025). B-CALM uses the empirical-Bayes average over σ_Δ. Pooled power is 0.935–1.000 in every cell.

	B-CALM				Pooled		Trial-only			
	Type I		Power		Type I		Type I		Power	
Bias	5k	20k	5k	20k	5k	20k	5k	20k	5k	20k
Shared-only										
None	0.010	0.010	0.745	0.765	0.030	0.035	0.025	0.020	0.600	0.620
Baseline only	0.010	0.005	0.745	0.755	0.065	0.055	0.025	0.020	0.600	0.620
Comp., $s_\Delta = 0.2$	0.055	0.055	0.835	0.855	1.000	1.000	0.030	0.015	0.610	0.650
Comp., $s_\Delta = 0.5$	0.090	0.105	0.780	0.745	1.000	1.000	0.030	0.030	0.640	0.605
Balanced										
None	0.000	0.005	0.785	0.825	0.015	0.000	0.025	0.015	0.585	0.615
Baseline only	0.000	0.005	0.780	0.800	0.015	0.000	0.025	0.015	0.585	0.615
Comp., $s_\Delta = 0.2$	0.060	0.060	0.800	0.805	1.000	1.000	0.030	0.015	0.580	0.620
Comp., $s_\Delta = 0.5$	0.090	0.110	0.860	0.860	1.000	1.000	0.030	0.025	0.630	0.605
Mismatch-heavy										
None	0.005	0.005	0.740	0.775	0.015	0.000	0.020	0.020	0.575	0.615
Baseline only	0.005	0.005	0.730	0.775	0.050	0.005	0.020	0.020	0.575	0.615
Comp., $s_\Delta = 0.2$	0.035	0.050	0.745	0.785	0.855	0.985	0.025	0.020	0.605	0.615
Comp., $s_\Delta = 0.5$	0.045	0.075	0.785	0.825	1.000	1.000	0.035	0.030	0.600	0.610

Table 13: Go/No-Go operating characteristics of B-CALM at each fixed comparative-bias scale σ_Δ (balanced design, $n^o = 5{,}000$, 200 replicates per cell). Each column pair reuses the component fits of the empirical-Bayes average; the $n^o = 20{,}000$ results are in the released code output and show the same pattern.

	No bias		Baseline only		Comp., $s_\Delta = 0.2$		Comp., $s_\Delta = 0.5$	
σ_Δ	Type I	Power	Type I	Power	Type I	Power	Type I	Power
0.01	0.010	1.000	0.010	1.000	0.995	1.000	1.000	1.000
0.05	0.000	0.940	0.000	0.940	0.445	1.000	1.000	1.000
0.2	0.005	0.645	0.005	0.640	0.035	0.795	0.110	0.915
0.5	0.015	0.605	0.015	0.590	0.030	0.630	0.035	0.670
1	0.020	0.590	0.020	0.580	0.030	0.585	0.030	0.630
∞	0.025	0.590	0.020	0.580	0.030	0.580	0.030	0.625

Table 14: Main factorial grid under three prior weight vectors on $\sigma_\Delta \in \{0.01, 0.05, 0.2, 0.5, 1.0, \infty\}$: the vector used in all experiments, $(0.01, 0.01, 0.42, 0.21, 0.13, 0.22)$; uniform weights; and a concentrated vector, $(0.01, 0.01, 0.65, 0.15, 0.10, 0.08)$. Columns give grid means of RMSE, coverage, width, and NT; the lowest cell coverage; and coverage under comparative bias of strength 0.2 and 0.5.

Prior weights	RMSE	Cov.	Width	NT	Min. cov.	Cov. (0.2)	Cov. (0.5)
Used	0.206	0.939	0.739	0.023	0.857	0.928	0.858
Uniform	0.179	0.893	0.571	0.052	0.731	0.784	0.849
Concentrated	0.210	0.937	0.751	0.022	0.853	0.932	0.853

E Posterior calibration and sensitivity plots

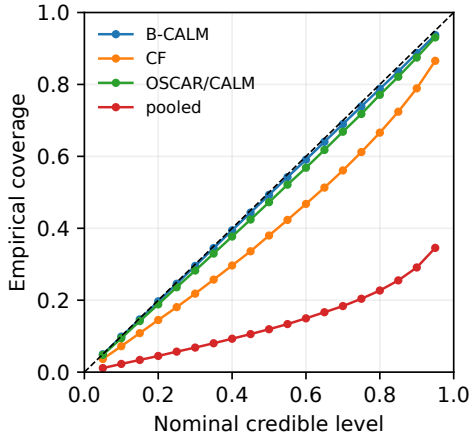


Figure 5: Posterior calibration on the balanced DGP under comparative bias: empirical coverage against nominal credible level. B-CALM tracks the diagonal while pooled and CF under-cover; the OSCAR/CALM family (R-OSCAR, MR-OSCAR, and CALM agree to within 10^{-4} at every level and are plotted as one line) sits just below nominal at the 90% mark.

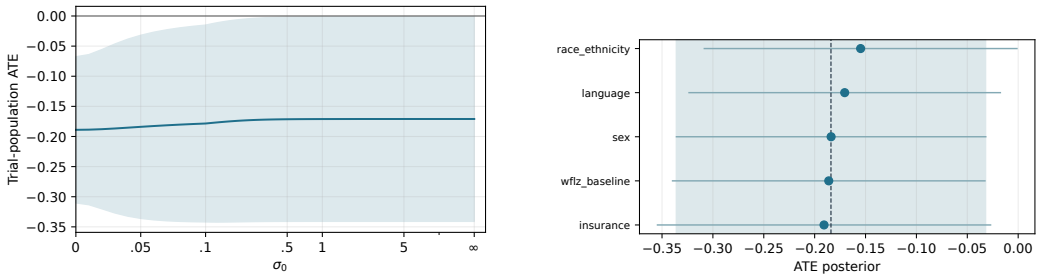
F Real-data application diagnostics

This section reports the diagnostics supporting the pediatric-obesity external-control analysis in Section 6. Table 15 gives the full real-data baseline comparison, including Bayesian causal forest and OSCAR-family baselines omitted from the compact main-text table. All B-CALM, trial-only, and pooled intervals in this section are exact Gaussian intervals; the trial-population ATE has posterior variance $\bar{\psi}^\top \Sigma_\tau \bar{\psi}$, where $\bar{\psi}$ is the average basis vector over the trial sample and Σ_τ the posterior covariance of the treatment-effect coefficients. Table 16 checks sensitivity to masking one EHR control covariate at a time. Tables 17 and 18 report where the EHR controls disagree most with the trial control surface and how much hidden-baseline-bias risk each method's interval would absorb under selected and skeptical priors. Figure 6 visualizes the σ_0 sensitivity curve and the leave-one-covariate diagnostics.

Bias-susceptibility ratio. The skeptical ratio in Table 15 is the 90% range of B-CALM-calibrated bias risk under a skeptical $\sigma_0 = 0.6$ prior (the default baseline-bias prior of the synthetic experiments, an order of magnitude above the selected 0.05), divided by a method's nominal width; values near or above one mean the interval is about as narrow as the bias it would absorb, or narrower. Pooling (0.98) and CF (2.91) achieve narrower intervals by accepting bias risk comparable to or larger than their reported width, while B-CALM at $\sigma_0 = 0.05$ stays well below one (0.32).

Table 15: Real-data external-control augmentation. ATE on the trial covariate distribution; width reduction is relative to RCT-only. B-CALM uses $\sigma_0 = 0.05$, selected over a σ_0 grid by a prior-weighted score (Appendix B.6); B-CALM, RCT-only, and pooled intervals are exact Gaussian intervals. With one-arm EHR controls, σ_Δ is held diffuse and σ_0 is the active borrowing knob. R-OSCAR, MR-OSCAR, and CALM agree to three decimals; Table 7 gives their estimates at full precision. The last column is the bias-susceptibility ratio under the skeptical $\sigma_0 = 0.6$ prior: 90% range of the implied B-CALM-calibrated bias risk on each method's ATE, divided by that method's nominal width. A value near or above 1 means a method's nominal interval is about as narrow as, or narrower than, the bias it would absorb under a skeptical EHR-control prior.

Method	ATE [90% CI]	Width (Red.)	Skep./w.
B-CALM	$-0.184\,[-0.336, -0.032]$	$0.304\,(10.5\%)$	0.32
RCT-only	$-0.171\,[-0.341, -0.001]$	0.340	0.00
pooled	$-0.189\,[-0.310, -0.068]$	$0.243\,(28.6\%)$	0.98
CF	$-0.162\,[-0.211, -0.112]$	$0.099\,(70.9\%)$	2.91
BCF	$-0.165\,[-0.242, -0.087]$	$0.155\,(54.3\%)$	–
R-OSCAR	$-0.178\,[-0.331, -0.005]$	$0.327\,(3.9\%)$	0.00
MR-OSCAR	$-0.178\,[-0.331, -0.005]$	$0.327\,(3.9\%)$	0.00
CALM	$-0.178\,[-0.331, -0.005]$	$0.327\,(3.9\%)$	0.00



(a) Trial-population ATE and exact 90% interval across σ_0.

(b) Leave-one-covariate masking.

Figure 6: Real-data external-control augmentation diagnostics. With one-arm EHR controls, σ_0 is the active borrowing knob and σ_Δ is held diffuse.

Table 16: Leave-one-covariate masking in the EHR controls. The final column checks whether the full-model 90% CI contains the masked posterior mean.

Mask	ATE	90% CI	Width	Hit
full	-0.184	[-0.336, -0.032]	0.304	–
sex	-0.184	[-0.336, -0.032]	0.304	✓
race_ethnicity	-0.155	[-0.308, -0.001]	0.307	✓
language	-0.170	[-0.323, -0.018]	0.306	✓
insurance	-0.191	[-0.354, -0.027]	0.327	✓
wflz_baseline	-0.186	[-0.340, -0.033]	0.307	✓

Table 17: Largest EHR-versus-RCT control-outcome discrepancies by posterior $|b_0(z)|$ under the skeptical commensurability prior $\sigma_0 = 0.6$. The skeptical prior gives b_0 enough room to absorb between-source baseline disagreement so that the diagnostic surfaces the strata where pooling would be most misleading; the trial-anchored ATE in Table 15 uses the score-selected $\sigma_0 = 0.05$ and yields much smaller posterior b_0 shrinkage.

Stratum	n	Mean	90% CI
Duke; No/other payer; base low	3027	-0.567	[-1.178, 0.044]
Duke; Medicaid; base low	542	-0.501	[-0.916, -0.086]
Duke; No/other payer; base mid	3022	-0.458	[-1.002, 0.086]

Table 18: B-CALM-calibrated hidden-bias susceptibility in the GPS external-control analysis. The uniform column is the ATE change under a $+0.1$ WFLz shift of all EHR control outcomes. Bias-risk ranges are posterior 5th–95th percentiles of $\sum_s L_{m,s} b_{0,s}$ or row-level analogues.

Method	ATE	Nominal 90% CI	Width	$+0.1$ shift	Selected risk	Sel./width	Skeptical risk	Skep./width
B-CALM	-0.184	[-0.336, -0.032]	0.304	-0.038	[-0.040, 0.033]	0.24	[-0.062, 0.036]	0.32
RCT-only	-0.171	[-0.341, -0.001]	0.340	0.000	[0.000, 0.000]	0.00	[0.000, 0.000]	0.00
pooled	-0.189	[-0.310, -0.068]	0.243	-0.096	[-0.097, 0.087]	0.75	[-0.136, 0.101]	0.98
CF	-0.162	[-0.211, -0.112]	0.099	-0.088	[-0.086, 0.115]	2.03	[-0.023, 0.265]	2.91
R-OSCAR	-0.178	[-0.331, -0.005]	0.327	0.000	[-0.000, 0.000]	0.00	[-0.000, 0.000]	0.00
MR-OSCAR	-0.178	[-0.331, -0.005]	0.327	0.000	[-0.000, 0.000]	0.00	[-0.000, 0.000]	0.00
CALM	-0.178	[-0.331, -0.005]	0.327	0.000	[-0.000, 0.000]	0.00	[-0.000, 0.000]	0.00
B-CALM $\sigma_0 = 0.6$	-0.171	[-0.341, -0.002]	0.339	-0.000	[-0.001, 0.000]	0.00	[-0.002, 0.001]	0.01
B-CALM $\sigma_0 = \infty$	-0.171	[-0.341, -0.001]	0.340	-0.000	[-0.000, 0.000]	0.00	[-0.000, 0.000]	0.00

G Extended related work

Multi-task learning. The outcome model of B-CALM has the same superposition as a standard device in multi-task learning, where each task's parameter is the sum of a component shared across tasks and a task-specific deviation [13]. The data shared lasso applies this device to sparse regression [15]. We studied its estimation theory in high dimension [3, 4] and used it to separate pan-cancer from tissue-specific drug-response signal [33, 34]. The two settings differ in their reference point. Multi-task learning treats the tasks symmetrically and learns them jointly, so the shared parameter is defined by the model. B-CALM has two sources, and the trial is the reference: its surface is the

estimand, while b_0 and b_Δ describe only how the OS departs from it. The prior on that departure, not the sparsity of a deviation, then limits how much the OS can inform the trial contrast.

Representation learning under treatment and domain mismatch. Representation-based CATE methods such as TARNet and CFRNet [23, 32] address treatment-arm imbalance within an observational study: they learn a representation $\Phi(X)$ and outcome heads for $Y(0)$ and $Y(1)$, with balance penalties or bounds involving an integral-probability metric between $P\{\Phi(X) \mid A = 1\}$ and $P\{\Phi(X) \mid A = 0\}$. HTCE [8] addresses a domain-transfer version of CATE estimation in which source and target domains may have different covariate spaces, using shared and private representations to transfer potential-outcome information. B-CALM uses representation learning for a different purpose: the mismatch is between randomized and observational sources, the estimand remains the RCT-defined CATE, and OS treatment-contrast information is allowed to help only through the explicit comparative-bias function $b_\Delta(z)$.

CATE uncertainty and RCT–OS fusion. Bayesian CATE learners provide posterior uncertainty for treatment-effect surfaces, including multi-task Gaussian processes for individualized effects [1] and Bayesian causal forests [16]. These methods model CATE uncertainty, but they do not by themselves solve the problem of borrowing from a potentially biased OS under covariate mismatch. A separate RCT–OS fusion literature uses trial data to correct or constrain observational CATE learning: Kallus et al. [25] use limited experimental data to correct hidden confounding in observational models, Hatt et al. [18] use observational data to learn shared representations and randomized data to learn trial-specific structure, and Ilse et al. [22] combine observational and interventional distributions through causal reductions. Closest to our Bayesian framing, Causal-ICM [12] uses multi-task Gaussian processes with a borrowing parameter for RCT–OS heterogeneous-treatment-effect estimation. B-CALM differs by targeting the RCT-defined CATE under partially overlapping covariates and by making source bias an explicit baseline and comparative-bias process whose prior controls local borrowing.

Relative to representation-based CATE and heterogeneous-feature transfer methods [8, 23, 32], the relevant shift is RCT-versus-OS rather than treated-versus-control, and the observational source is a biased proxy for the trial estimand rather than its own target population. Generalizing from the randomized-trial population to a broader target population is a distinct transportability problem, reviewed by Colnet et al. [10] and Degtiar and Rose [11]. More recently, it has been studied under outcome-distribution shift using several sensitivity-analysis frameworks, including the marginal sensitivity model [6] and the omitted-variable-bias framework [7].

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 Section 1 state the bias-limited borrowing claim, the RCT-defined CATE target, and the main theoretical and empirical contributions. Section 7 states the practical reporting target and the main limitations.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: Section 7 discusses non-identifiability of comparative bias without prior restrictions, non-uniform calibration and false-positive inflation of a Go/No-Go rule under comparative bias, and the efficiency cost relative to a scalar discrepancy when the bias is constant. Sections 5 and 6.3 state the fixed-representation scope of the information bounds and the baseline-only scope of the real-data application. Appendix A.11 expands on directions beyond the present scope.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes].

Justification: Section 5 states the assumptions for the scalar, function-valued, and PAC-Bayesian-style results. Complete derivations and proof details are provided in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes].

Justification: Section 6 and Appendix A.12 disclose the main designs, sample sizes, seed counts, metrics, and baselines; Appendix B gives the data-generating equations, the estimator and selection procedure as implemented, and the metric formulas. The public code repository contains the scripts, seeds, configurations, and commands that regenerate every synthetic and semi-synthetic result. The real-data analysis code is also in the repository; its input files and the patient-level data are not released.

Guidelines:

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

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes].

   Justification: Code, seeds, configurations, and reproduction instructions are available at https://github.com/AsiaeeLab/b-calm (Appendix A.12). Patient-level real-data records cannot be publicly released under the trial's data-use terms; sharing of the trial data is handled by the trial investigators through the funder.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: The paper specifies benchmark designs, grids, seeds, baselines, and key model settings in Section 6 and Appendix A.12, with the full specifications in Appendix B. All fits are closed form, so no optimizer settings are involved; the software environment is listed in the code repository.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes].

   Justification: The simulation tables report averages over 50 seeds and RMSE variability as across-seed standard deviations, and the real-data analysis reports posterior 90% credible intervals. We do not use formal null-hypothesis significance tests.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: Appendix A.12 states that the reported experiments were run on a single workstation (Intel Xeon w9-3575X, no GPU), using up to about ten parallel worker processes. The main factorial grid takes several hours with the bootstrapped baselines and about two minutes for B-CALM alone; the calibration study takes about 25 minutes, and every other experiment takes minutes or less.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: To the best of our knowledge, the work conforms to the NeurIPS Code of Ethics. The real-data analysis uses governed patient data and does not release patient-level records.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes].

Justification: Appendix A.10 discusses the intended positive impact of cautious external-control augmentation and the negative risk of overinterpreting biased observational evidence.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A].

Justification: The paper does not release high-risk pretrained models, generative systems, or scraped datasets. Patient-level clinical data are not released.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes].

Justification: We cite the methodological baselines and data sources used in the experiments. Code assets and licenses: EconML (MIT), PyMC and PyMC-BART (Apache 2.0), scikit-learn (BSD-3-Clause), and the QR-learner implementation of Karlsson et al. [26] (MIT), redistributed unmodified with its license notice. The real-data analysis uses the Greenlight Plus trial and linked EHR data under the source studies' data-use terms (Appendix A.10).

Guidelines:

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

13. **New assets**

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

    Answer: [Yes].

    Justification: The released code repository (MIT license) documents installation, the reproduction commands, and the mapping from scripts to the paper's tables and figures. No new datasets are released.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The work does not involve crowdsourcing or newly recruited human-subject experiments. The real-data analysis uses previously collected data under source-study governance.

    Guidelines:

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

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

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

Answer: [Yes].

Justification: Appendix A.10 states that the real-data analysis uses previously collected data under source-study data-use and institutional-review constraints, and Appendix A.12 states that patient-level data are not released.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A].

Justification: LLMs are not part of the proposed method, experiments, or theoretical results.

Guidelines:

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