Computationally Efficient Estimation of Localized Treatment Effects for Multi-Level, Multi-Component Interventions to Address the Opioid Crisis

Abdulrahman A. Ahmed[a] , M. Amin Rahimian *,[a] , Qiushi Chen [b], and Praveen Kumar[c]

[a] Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, USA;
[b] Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, USA;
[c] Department of Health Policy and Management, University of Pittsburgh, Pittsburgh, USA

ABSTRACT

The opioid epidemic remains a major public health challenge in the United States, requiring a *multi-pronged intervention* approach to mitigate harms to communities. Given the heterogeneity of the epidemic across the country, it is crucial for policymakers to understand *localized treatment effects* of different intervention components and utilize limited resources efficiently. While locally calibrated simulation models offer a useful computational tool to project the epidemic outcomes for any given intervention policy, collecting simulation results for all intervention combinations to estimate localized treatment effects for each community is impractical because the number of possible intervention combinations grows exponentially with the number of interventions and levels at which they are applied. To tackle this, we develop a *bi-level* metamodel framework with a two-stage sequential design for efficient sampling. The metamodel consists of a response function linking health outcomes to each intervention component's treatment effect, and a Gaussian process regression (GPR) to learn spatial and socio-economic structures of the treatment effects based on locally-contextualized covariates. With two-stage sequential sampling, we leverage spatial correlations and posterior uncertainty to sequentially sample the most informative counties and treatment conditions. We apply this framework to estimate treatment effects of buprenorphine dispensing and naloxone distribution on overdose mortality rates using a calibrated agent-based opioid epidemic model in Pennsylvania counties. Our approach achieves approximately 5% average relative error using one-tenth the number of runs required for an exhaustive simulation. Our bi-level framework provides a computationally efficient approach to support policymakers, enabling an efficient evaluation of alternative resource-allocation strategies to mitigate the opioid epidemic in local communities.

KEYWORDS
metamodel; large-scale simulation; sequential design; model selection; epidemiological models; precision public health

1. Introduction

The opioid crisis remains a major public health challenge, with more than 750,000 deaths in the United States over the past two decades (Garnett & Miniño, 2024). A

*Correspondence to: rahimian@pitt.edu

defining feature of this epidemic is its pronounced heterogeneity across the country. For example, the drug overdose mortality rates in 2023 ranged from 9.0 per 100,000 in Nebraska to 81.9 per 100,000 in West Virginia (Centers for Disease Control and Prevention, 2025). Moreover, the epidemic has been evolving constantly, shifting from prescription opioids to heroin, then to synthetic opioids like fentanyl, and more recently to co-stimulant use as the primary drivers of mortality (Volkow & Blanco, 2021; Jenkins, 2021; Ciccarone, 2021; Ahmed et al., 2022). These complex dynamic and geospatial patterns have contributed to the variation seen in the extent of the opioid crisis across states and counties in the US.

It has become increasingly clear that a multi-pronged intervention approach is needed to mitigate the harm by the opioid crisis (Blanco et al., 2020). It requires a coordinated set of interventions spanning prevention, treatment and harm reduction strategies. These interventions include implementing policies that support non-opioid pain management therapies, enhancing prescriber education, reducing stigma through public education campaigns, and improving access to medications for opioid use disorder (MOUD) by removing barriers to treatment (H. Cheng et al., 2022). Additional efforts have focused on increasing initiation and retention in care, facilitating linkage to care with peer support, and expanding access to fentanyl test strips and naloxone kits (D'Onofrio et al., 2015; Liebschutz et al., 2014). Among the most widely deployed and evidence-based interventions available to county health departments are naloxone distribution, a harm-reduction strategy that reverses opioid overdoses, and buprenorphine, a medication for opioid use disorder (MOUD) that reduces disease burden by supporting recovery (Cerdá et al., 2021; Lim et al., 2022).

Critically, the effectiveness of any given intervention combination is not uniform across communities. Counties differ substantially in their epidemic trajectory, demographic composition, and urban or rural character, such that the same multi-pronged strategy may produce markedly different reductions in overdose mortality depending on local conditions (Cerdá et al., 2024; Dodson, Enki Yoo, Martin-Gill, & Roth, 2018; Marotta et al., 2019; Zheng et al., 2025). With the wide range of evidence-based interventions available but limited resources, policymakers need to assess the priority of available interventions and develop an effective intervention package that consists of multiple intervention components with each at a proper level. More importantly, given the heterogeneity of the crisis and the varying stages of the epidemic across counties, the multi-level, multi-component interventions must be tailored to meet the unmet needs of local communities, such as counties. A key to facilitating such decision making process is to understand the localized treatment effects for each intervention component. That is, impact of the same level of intervention may not necessarily translate into similar impact on opioid and substance-related outcomes, such as overdose deaths.

While several modeling-based studies have utilized simulation to project the impact of various intervention strategies (Scheidell et al., 2024; Irvine et al., 2022; Lim et al., 2022; Zang et al., 2022), estimating the impact of multi-level, multi-component interventions at local level is a non-trivial task, even with a validated simulation model. A naive strategy is to exhaustively enumerate and evaluate all possible combinations of interventions for a local county, estimate the treatment effect in this county, and then repeat for each county. Clearly, such full factorial design leads to a rapid expansion in the total number of design points to be evaluated (by simulation). If each of J interventions can be assigned ℓ levels, the total number of treatment combinations grows exponentially as ℓ^J. To estimate treatment effects for different counties, the total number of required simulation runs further increases multiplicatively. For example,

with 50 counties and five interventions at five levels each, a full factorial design entails $50 \times 5^5 = 156{,}250$ configurations, and this count increases further when multiple replications are needed for stochastic simulations, posing a significant computational challenge.

While in practice, fractional factorial, Latin hypercube, or other space-filling designs are commonly used to mitigate this burden (Sanchez et al., 2020), these approaches still face a trade-off: one can capture higher-order interactions and policy synergies, but only with substantial computational cost; or reduce the number of runs, but at the price of overlooking some of these effects. These trade-offs, combined with the spatial correlation across counties, the stochastic variability of epidemic simulation outputs, and the exponentially growing design space, make large-scale simulation modeling particularly challenging.

Metamodels, or surrogate models, offer a solution by learning statistical mappings from inputs to simulation outputs. Once trained, a metamodel can provide rapid predictions and quantify uncertainty for input parameter that were not directly simulated, thereby enabling efficient design evaluation and sensitivity analysis. Among various metamodeling techniques, Gaussian process regression (GPR) is widely used because it flexibly captures nonlinear functions and produces calibrated uncertainty estimates (Rasmussen & Williams, 2006; Forrester et al., 2008; Gramacy, 2020). The foundational work of Kennedy & O'Hagan, 2001 and subsequent extensions by Conti & O'Hagan, 2010 demonstrate how GPR can emulate expensive computer models while propagating parameter uncertainty. To improve learning efficiency, metamodeling is often coupled with sequential design, which adaptively selects the most informative simulation runs based on current model predictions, focusing resources on the most uncertain regions of the input space (Frazier, 2018; Wilson et al., 2018; Balandat et al., 2020). Acquisition functions such as predictive variance, entropy reduction, or expected improvement guide this selection and are particularly useful when the number of input combinations is large or when simulation cost is high (Fisher et al., 2020).

Metamodeling, including GPR, has been applied to opioid epidemic settings, with some specifically focusing on estimating treatment effect of intervention policy. Ahmed, Rahimian, & Roberts, 2023a propose a regression-based greedy sampling strategy that allocates simulation effort across treatment conditions based on confidence interval width, achieving accuracy comparable to uniform sampling with substantially fewer simulations. However, their work is limited to a single county and cannot capture the spatial and socio-economic structures to account for that induce county-level heterogeneity that is crucial for understanding the opioid epidemic.

This spatial structure arises naturally in opioid simulation models, where outputs are summarized as coefficients describing how each treatment affects overdose death rates. For a single county, this can be achieved using regression-based estimates. For example, Ahmed, Rahimian, & Roberts, 2024 study the optimal selection of linear regression functions to estimate treatment effects given a limited number of simulations for a single county. However, on the larger geographic scales of entire states with many counties, these coefficients become interdependent and geographically structured: counties that are spatially proximate or demographically similar frequently share treatment-response patterns (Banerjee et al., 2003). Ignoring this structure can lead to inefficient sampling and incoherent predictions.

Despite the increasing use of GPR in public health applications, current practice mainly focuses on prediction, calibration, or spatial risk mapping in limited policy spaces (Senanayake et al., 2016; Zimmer & Yaesoubi, 2020a). No existing framework integrates spatially-aware GPR with a sequential design that scales to large policy

spaces, which is needed to address the challenge of estimating treatment effects of multi-level, multi-component opioid intervention policies across counties. In this paper, We propose a bi-level metamodel which consists of the GPR as the first level to model contextual variability of treatment effects across counties, and a response function as the second level to map the GPR predictions to overdose mortality outcome. We design a two-stage sequential design that allocates simulation runs across counties and the intervention combinations to maximize information gain.

1.1. Related Work

GPR is commonly used in epidemiological settings and for emulating computationally intensive simulations in public health, as illustrated in the following paragraphs; yet its application to opioid epidemic remains largely unexplored.

For example, Langmüller et al., 2024 develop GPR surrogates to emulate dengue outbreak simulations across an eight-dimensional parameter space, enabling efficient evaluation of outbreak probability and epidemic duration. Similarly, Sawe et al., 2024 employ GPR to approximate multi-disease agent-based simulations in Kenya, reducing computation time more than ten-fold while preserving predictive accuracy. In the context of malaria, Reiker et al., 2021 use GPR-based Bayesian optimization to calibrate high-dimensional transmission models, highlighting its value for accelerating policy-relevant inference. Influenza forecasting studies further illustrate the utility of GPR for spatio-temporal epidemic prediction (Senanayake et al., 2016; Zimmer & Yaesoubi, 2020b). In more recent work, Ahmed, Rahimian, & Roberts, 2023b demonstrate the potential of GPR for capturing differences in spatial distribution of outcomes for different diseases, which can be attributed to differences in their underlying epidemic dynamics, when other confounding factors such as population size, location, and contact rates are kept identical in simulations using synthetic populations. However, their approach relies on fitting a single-output GPR surrogate to a single county's population and is not applicable for evaluating the heterogeneous effects of multi-component interventions in different counties.

A parallel stream of work applies spatial GPR to disease risk mapping and inference. Classic geostatistical methods (Banerjee et al., 2003; Moraga, 2023) and large-scale malaria risk maps (Bhatt et al., 2017) show the strength of spatial kernels for interpolating across heterogeneous regions. However, these studies primarily focus on observational prediction and mapping rather than emulating county-level policy simulations. Spatial structure has rarely been integrated with surrogate modeling for exponentially large policy spaces. Appendix A provides more detailed related work on the application of GPR in epidemiological modeling and public health.

Another relevant strand of work centers on healthcare and biomedical applications of GPR to provide flexible function approximation and uncertainty quantification. For example, GPR has been applied to ICU monitoring (L.-F. Cheng et al., 2020), pharmacology, and to predict dose-response curves (Gutierrez et al., 2024). These studies demonstrate the advantages of modeling correlated outputs but do not address spatial heterogeneity or simulation-based policy learning.

Additionally, sequential design and active learning methods are well established for simulation emulation and Bayesian optimization (Frazier, 2018; Wilson et al., 2018). Approaches such as expected improvement, predictive variance, and entropy reduction have been applied broadly, and recent work explores active learning with multi-output surrogates (Li et al., 2022). Yet, in healthcare applications, these methods are generally

used for calibration or parameter tuning, not for the hierarchical problem of allocating simulation runs across counties and treatment conditions.

In summary, existing literature demonstrates the utility of GPR for epidemic simulation, spatial prediction, and correlated outputs. However, no framework to date integrates spatially-aware GPR with a hierarchical sequential design that efficiently explores exponentially many opioid intervention combinations across heterogeneous counties. Our work addresses this gap by developing a bi-level metamodeling framework that combines GPR with a response function and introduces a two-stage sequential design for allocating simulations across both counties and interventions.

1.2. Main Contributions

In this work, we develop a novel, sample-efficient metamodeling framework for estimating opioid intervention effects across multiple counties, that integrates spatial GPR, linear outcome models, and a two-stage sequential design strategy for selective simulation. While our analysis and results are tailored to the opioid epidemic in Pennsylvania, the proposed analytical framework readily generalizes to other states and intervention modalities. Our main methodological contributions are as follows:

(1) GPR-based modeling of spatially varying response-function coefficients. We extend traditional GPR metamodeling to a spatially-structured setting that captures demographic heterogeneity and geographic correlation across counties. Each county is encoded by its centroid coordinates and a set of socio-economic features, such as income, racial composition, and population density, forming a high-dimensional input space. The model outputs county-specific *coefficients* for a response function, representing the *treatment effect*, which translates treatment levels into predicted overdose death rates. Each coefficient of the response function is modeled by its own Gaussian process defined over the same spatial and socio-economic input space, producing cross-sectional mortality predictions for a target time point (e.g., at the end of a five-year period over which interventions are planned).

In addition, we incorporate a heteroscedastic noise model in which the observation variance of each coefficient is estimated from the sample variance and the number of simulation replicates selected for each county. This formulation naturally captures county-specific uncertainty, arising from differences in population size, urban–rural structure, and other socio-economic factors, and allows the metamodel to weight observations according to their estimated precision, driven by the number of simulation replicates and the resulting regression-coefficient variance at each county, yielding more reliable posterior estimates than a homoscedastic specification. Our kernel design combines multiple radial basis function kernels, allowing the model to represent smooth spatial variation across regions.

(2) Two-stage sequential design for efficient sampling. To efficiently sample the input space of multi-component interventions across multiple counties, we develop a two-stage sequential design procedure: In the first stage, we select which counties to simulate based on their posterior uncertainty in the GPR model. Specifically, we adapt the Signal-to-Noise Ratio, defined as the ratio of posterior standard deviation to posterior mean, as an acquisition function to prioritize counties with the most uncertain model estimates. In the second stage, for the selected county, we choose the treatment condition whose predicted outcome has the most posterior uncertainty, as measured by the width of its credible interval. This sequential design enables us to efficiently target simulations to regions where they are most needed, thereby accelerating convergence

while maintaining model accuracy.

The remainder of the paper is organized as follows. Section 2 introduces the bi-level metamodel that uses Gaussian process regression to estimate the coefficients of a response function and learn their contextual dependencies on county-specific features. Section 3 describes the two-stage sequential design framework, detailing how county and treatment-condition sampling are integrated under a unified bi-level procedure. Section 4 presents empirical results on model performance, kernel design, and sample complexity, including learning curves and estimated treatment effects using a calibrated model of opioid use disorder for counties in Pennsylvania. Finally, Section 5 concludes with a summary of our findings, implications for policy evaluation, and directions for future work.

2. Problem Formulation and Modeling Framework

Policymakers responding to the opioid epidemic face a fundamental challenge: identifying which intervention combinations work best for each community. This goal, tailoring public health strategies to local conditions rather than applying uniform policies, reflects the emerging paradigm of precision public health. Achieving it requires evaluating how different communities respond to different combinations of interventions such as naloxone distribution and buprenorphine treatment access.

A brute-force approach would calibrate a simulation model for each county independently and exhaustively evaluate all possible policy combinations. This is computationally inefficient. In our setting, each county is subject to five naloxone levels and five buprenorphine levels, yielding $5 \times 5 = 25$ distinct treatment conditions. Each condition requires hundreds of simulation replicates to reduce variability, and with 67 counties, exhaustive evaluation would demand millions of runs. The problem intensifies as the intervention space expands: six interventions at seven levels each would yield 7^6 combinations. This challenge generalizes beyond opioid modeling. Any setting involving simulation-based policy evaluation across heterogeneous subgroups (counties, demographic strata, healthcare facilities) with multi-dimensional treatment spaces faces the same combinatorial barrier. The subgroups differ in baseline characteristics, the interventions operate at multiple levels, and the treatment response varies across both subgroups and intervention combinations.

Our goal is to develop a modeling framework that efficiently estimates subgroup-level treatment effects across a high-dimensional intervention space without exhaustively simulating every subgroup-treatment combination. Specifically, the framework must: (1) generalize across subgroups by learning how baseline characteristics shape treatment response, (2) interpolate across treatment levels to predict outcomes for less simulated and unsimulated conditions, and (3) quantify predictive uncertainty to guide sequential allocation of simulation effort.

2.1. Bi-level Metamodel: GPR and Response Function Modeling

A conventional approach to metamodeling would employ a GPR to map county features directly to outcomes across all treatment conditions. This approach requires learning a function that maps county-level features to outcomes for all ℓ^J treatment conditions simultaneously, which suffers from the curse of dimensionality and is highly data inefficient.

Figure 1 provides a schematic overview of our proposed bi-level metamodel. Instead

of learning outcomes separately for every treatment combination, the framework uses GPR to learn county-specific response-function coefficients, which are then used to compute outcomes for any treatment level through a response function. Specifically, for a given county c and treatment condition (n, b), where n and b take integer values in $\{1, 2, \ldots, \ell\}$ and encode naloxone and buprenorphine levels, we model the opioid overdose death rate, measured as deaths per 100,000 people, using the following linear response function:

$$z(n, b \mid c) = \mu_0(\mathbf{x}_c) + \mu_n(\mathbf{x}_c) \cdot n + \mu_b(\mathbf{x}_c) \cdot b. \tag{1}$$

In our bi-level metamodel framework, we refer to the response function $z(n, b \mid c)$ as the outcome level, which maps treatment condition (n, b) to the outcome of interest (predicted overdose mortality) for each county c. We adopt a main-effects specification after examining factorial plots, which indicate no interaction between naloxone and buprenorphine across counties; details are provided in Appendix B. In this formulation, $\mathbf{x}_c$ is a feature vector of spatial and socio-economic covariates for county c. For each county c, the coefficient vector $\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c),\ \mu_n(\mathbf{x}_c),\ \mu_b(\mathbf{x}_c)]^\top$ in (1) is learned as the posterior mean functions of three GPRs evaluated at $\mathbf{x}_c$.

Rather than modeling this outcome separately for every treatment condition, we express it through a parametric response function whose coefficients vary across counties. These coefficients are learned using a GPR model defined over spatial and socio-economic county features. This construction yields a bi-level metamodel: a contextual level, which models how response-function parameters depend on county characteristics, and an outcome level, which maps treatment levels (n, b) to predicted overdose mortality using the response function.

2.2. Contextual Modeling of Subgroup Heterogeneity using GPR

In the first level, we use GPR to learn the contextual dependencies of μ_0, μ_n, and μ_b on county-specific spatial and socio-economic features $\mathbf{x}_c$ in Equation (1). Specifically, to each coefficient $\mu_m, m \in \{0, n, b\}$ of the response function we associate a Gaussian process $\mathcal{GP}(\mu_m(\cdot), k_m(\cdot, \cdot))$, whose mean function evaluated at $\mathbf{x}_c$ gives the coefficient value for county c. The kernel function $k_m(\cdot, \cdot)$ determines the variance-covariance relations between county estimates based on their spatial and socio-economic features. Common kernels such as radial basis function have a width hyperparameter that controls similarities between county responses based on their covariates and is optimized separately (using maximum likelihood or other fit criteria). More complex kernels can be constructed as a composition of simpler kernels, for example, through addition to capture independent effects or multiplication to encode interactions among input features (Rasmussen & Williams, 2006, Chapter 4). In our implementation, the kernel encodes spatial and demographic similarity via a combination of multiple radial basis functions, defined in Appendix Equations (B1) and (B2). Further discussion of the kernel design for our study is provided in Appendix B.

GPR uses Bayes' rule to yield a posterior distribution over the response-function coefficients. The posterior provides both mean estimates and uncertainty for each coefficient, with uncertainty contracting as additional simulation runs are collected for a county. These posterior summaries are subsequently propagated through the outcome-level response function to generate uncertainty-aware predictions of overdose mortality.

Specifically, For a new county $\mathbf{x}_{c^*}$, the posterior mean and variance are given by

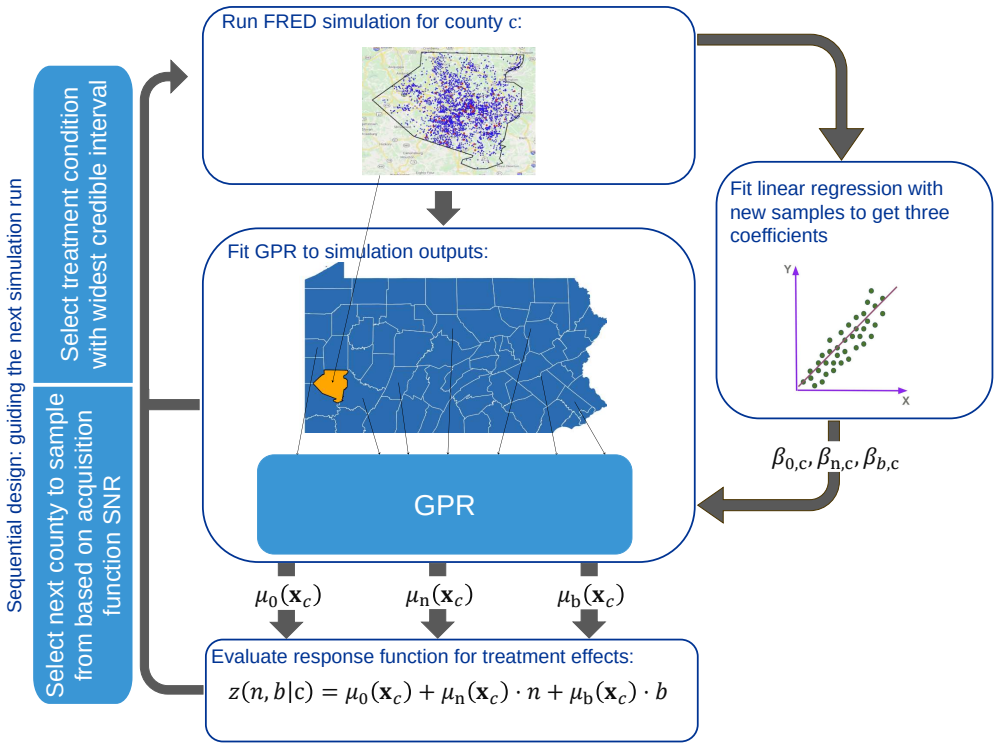


Figure 1: An overview of the proposed metamodeling framework for evaluating opioid interventions across Pennsylvania counties.
Top: County-level opioid overdose death outcomes are generated in the FRED simulation platform. The "FRED: Framework for Reconstructing Epidemiological Dynamics" is an agent-based simulation platform which is described in Appendix C, along with our opioid use disorder (OUD) model.
Middle: Three GPRs are fit to model three regression coefficients of the response function that estimates overdose death rates for different treatment combinations across all counties.
Bottom: The fitted GPR provides posterior distributions over the parameters of the linear response function, which is then used to evaluate treatment effects under any combination of naloxone (n) and buprenorphine (b) levels. Sequential design guides the most informative selection of counties and treatment conditions for subsequent simulation runs.

$\mu_m(\mathbf{x}_{c^*}) = \mathbf{k}_{c^*}^\top(\mathbf{K} + \mathbf{\Sigma}(\mathbf{x}_{c^*}))^{-1}\mathbf{y}$ and $\sigma_m^2(\mathbf{x}_{c^*}) = k(\mathbf{x}_{c^*}, \mathbf{x}_{c^*}) - \mathbf{k}_{c^*}^\top(\mathbf{K} + \mathbf{\Sigma}(\mathbf{x}_{c^*}))^{-1}\mathbf{k}_{c^*}$, where $\mathbf{K}$ is the kernel matrix evaluated at training counties, $\mathbf{k}_{c^*}$ is the vector of covariances between training counties and $\mathbf{x}_{c^*}$, $\mathbf{y}$ is the vector of observed coefficient values across sampled counties, and $\mathbf{\Sigma}(\mathbf{x}_{c^*}) = \mathrm{diag}([\sigma_0^2(\mathbf{x}_c), \sigma_n^2(\mathbf{x}_c), \sigma_b^2(\mathbf{x}_c)]^\top)$ is the heteroscedastic noise matrix. As a new sample is collected, the posterior is updated by augmenting the training set and recomputing the above expressions, with hyperparameters re-optimized using marginal likelihood maximization (Rasmussen & Williams, 2006, Chapter 2).

To train the Gaussian process (GP) using Bayes' rule, simulation outputs are converted into regression-based observations of the response-function coefficients. For a fixed county c, simulation outcomes evaluated at selected treatment conditions (n, b) are summarized using a linear regression,

$$r(n, b \mid c) \;=\; \beta_{0,c} + \beta_{n,c} \cdot n + \beta_{b,c} \cdot b, \tag{2}$$

where $\beta_{0,c}$, $\beta_{n,c}$, and $\beta_{b,c}$ are county-specific regression coefficients. Each set of simulation runs for county c therefore yields updated estimates of these coefficients.

Given the regression-based coefficient estimates as observations, GP training involves two distinct steps. First, the kernel hyperparameters $\boldsymbol{\theta}$ are optimized by maximizing the marginal log-likelihood,

$$\log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \boldsymbol{\Sigma}(\mathbf{x}_c))^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \boldsymbol{\Sigma}(\mathbf{x}_c)| - \frac{n}{2}\log 2\pi,$$

via L-BFGS through `BoTorch`'s `fit_gpytorch_mll` routine, and this re-optimization is performed at every sequential design iteration as new county–condition observations are added to the training set. Second, given the optimized hyperparameters, the GPR posterior over the response-function coefficients is updated analytically via Bayes' rule. For each county c with feature vector $\mathbf{x}_c \in \mathbb{R}^d$, the posterior mean $\mu_m(\mathbf{x}_c)$ and variance $\sigma_m^2(\mathbf{x}_c)$ for each coefficient $m \in \{0, n, b\}$ are given by $\mu_m(\mathbf{x}_c) = \mathbf{k}_c^\top(\mathbf{K}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma})^{-1}\mathbf{y}_m$ and $\sigma_m^2(\mathbf{x}_c) = k(\mathbf{x}_c, \mathbf{x}_c) - \mathbf{k}_c^\top(\mathbf{K} + \boldsymbol{\Sigma}(\mathbf{x}_c))^{-1}\mathbf{k}_c$, where $\mathbf{K}$ is the kernel matrix evaluated over all training counties, $\boldsymbol{\Sigma}(\mathbf{x}_c)$ is the diagonal heteroscedastic noise matrix derived from the regression variances, $\mathbf{k}_c$ is the vector of kernel evaluations between $\mathbf{x}_c$ and all training points, and $\mathbf{y}_m$ is the vector of observed regression coefficients for output m. Together, these posterior means form the updated coefficient vector $\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c),\, \mu_n(\mathbf{x}_c),\, \mu_b(\mathbf{x}_c)]^\top$.

We adopt an independent-output formulation, in which each output component $\mu_m(\mathbf{x}_c)$ is modeled by its GP over the same d-dimensional spatial–socioeconomic feature space $\mathbf{x}_c \in \mathbb{R}^d$. This choice simplifies the implementation while remaining well-justified: the GP models the coefficients $\beta_0, \beta_n, \beta_b$ of a linear regression refit at each iteration from all observed simulation runs for the selected county, so the coefficients are inherently correlated through the joint regression fit and further combined jointly through the response function (1) at the outcome level. Consequently, although the three GPs are fit independently at the contextual level, the sequential design operates on joint outcome predictions rather than individual coefficients in isolation. Empirical analysis of pairwise correlations among iterative coefficient updates is provided in Section 4.

The regression coefficients obtained from simulation runs are modeled as noisy observations of the GP outputs. For county c, the observation model for estimating coefficient $\mu_m(\mathbf{x}_c)$ is $y \mid f(\mathbf{x}_c) \sim \mathcal{N}\big(f(\mathbf{x}_c), \sigma_m^2(\mathbf{x}_c)\big)$, where $f(\mathbf{x}_c)$ is the output drawn from $\mathcal{GP}(\mu_m(\cdot), k_m(\cdot,\cdot))$ at point $\mathbf{x}_c \in \mathbb{R}^d$. The output noise $\sigma_m^2(\mathbf{x}_c)$ models the heteroscedastic uncertainty arising from finite simulation replicates and the variability of the regression-based estimates obtained from Equation (2). Specifically, for each county c and coefficient $\mu_m(\mathbf{x}_c)$, the noise variance is set equal to the estimated variance of the corresponding regression coefficient, $\sigma_m^2(\mathbf{x}_c) \propto \mathrm{Var}(\widehat{\beta}_{m,c}) \,/\, R_c$, where $\widehat{\beta}_{m,c}$ is the estimated regression coefficient for output m in county c, $\mathrm{Var}(\widehat{\beta}_{m,c})$ is its regression-based variance, and R_c denotes the number of simulation replicates selected for county c.

As additional replicates are collected, $\sigma_m^2(\mathbf{x}_c)$ decreases, enabling the GPR to represent heterogeneity across counties in a variance-aware manner, rather than treating all counties as equally informative. In contrast to a homoscedastic specification, which assumes a constant noise level across counties and coefficients, the heteroscedastic formulation adopted here explicitly links observation noise to the number of simulation replicates used to estimate each regression coefficient. As a result, uncertainty from the regression stage propagates coherently through the GPR and into the final

outcome-level predictions.

Given that these simulations are computationally expensive, how can we efficiently estimate effects across high-dimensional spaces under a limited simulation budget? We address this challenge by adopting a sequential design approach that leverages model uncertainty to guide sampling decisions. The next section formalizes this into a two-stage sequential design that jointly allocates simulation effort across both counties and treatment conditions.

3. Two-Stage Sequential Design for Joint Sampling of Counties and Treatment Conditions

In large-scale simulation settings, where evaluating every input configuration is computationally prohibitive, sequential design strategies enable efficient learning by guiding the sampling process to the most informative regions of the input space. Our approach consists of a two-stage sequential design framework: first, selecting which counties to sample from, and second, choosing which treatment conditions to evaluate for those counties. Both stages are guided by posterior uncertainty derived from the metamodel.

3.1. First-Stage Sequential Design for Sampling Counties

To allocate simulation resources efficiently, we prioritize sampling from counties where the metamodel exhibits high predictive uncertainty relative to the expected outcome. The GPR kernel guides both the similarity structure and uncertainty estimation, thereby influencing the sequential sampling trajectory for counties. We use the signal-to-noise ratio (SNR) as our acquisition function to determine which counties to sample from. This strategy is designed to improve global model accuracy across a high-dimensional spatial domain by focusing computational resources on regions where the model is least certain relative to the scale of the predicted effect.

Acquisition function formulation. For a given county c, the GPR posterior yields estimates of the response-function parameters, consisting of the posterior mean vector $\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c), \mu_n(\mathbf{x}_c), \mu_b(\mathbf{x}_c)]^\top$ and the associated diagonal matrix of observation noise variances $\boldsymbol{\Sigma}(\mathbf{x}_c) = \mathrm{diag}([\sigma_0^2(\mathbf{x}_c), \sigma_n^2(\mathbf{x}_c), \sigma_b^2(\mathbf{x}_c)]^\top)$. In our sequential design framework, the acquisition function is calculated directly from the posterior mean and covariance.

To compute a single acquisition value from these three posteriors, we employ a scalarized posterior transform. This transform applies a fixed weight vector $\mathbf{w} = [1/3, 1/3, 1/3]^\top$ to the posterior distribution, effectively averaging the predictions across all three parameters:

$$\mu = \mathbf{w}^\top \boldsymbol{\mu}(\mathbf{x}_c) = \frac{1}{3}\mu_0(\mathbf{x}_c) + \frac{1}{3}\mu_n(\mathbf{x}_c) + \frac{1}{3}\mu_b(\mathbf{x}_c), \quad \sigma^2 = \mathbf{w}^\top \boldsymbol{\Sigma}(\mathbf{x}_c)\mathbf{w},$$

where μ and σ^2 are the combined mean and variance. More generally, $\mathbf{w}$ may be specified to emphasize particular components of the response function, depending on modeling objectives or policy focus. The SNR acquisition function is then defined as:

$$\alpha_{\mathrm{SNR}}(c) = \frac{\sigma}{\mu} = \frac{\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma}(\mathbf{x}_c)\mathbf{w}}}{\mathbf{w}^\top \boldsymbol{\mu}(\mathbf{x}_c)}.$$

Sequential sampling procedure for counties. At iteration t, the GPR posterior for county c provides $\boldsymbol{\mu}_t(\mathbf{x}_c)$ and $\boldsymbol{\Sigma}_t(\mathbf{x}_c)$, where $\boldsymbol{\mu}_t(\mathbf{x}_c)$ and $\boldsymbol{\Sigma}_t(\mathbf{x}_c)$ denote the mean vector and the diagonal noise covariance matrix evaluated at the county feature vector $\mathbf{x}_c$ after incorporating all data available up to iteration t. These quantities are scalarized as $\mu_t = \mathbf{w}^\top \boldsymbol{\mu}_t(\mathbf{x}_c)$ and $\sigma_t = (\mathbf{w}^\top \boldsymbol{\Sigma}_t(\mathbf{x}_c)\mathbf{w})^{1/2}$. The SNR acquisition $\alpha_{\text{SNR}}(c) = \sigma_t/\mu_t$ selects the next county using $c^* = \arg\max_{c\in\mathcal{C}} \alpha_{\text{SNR}}(c)$. The first stage sequential design steps are summarized below:

Algorithm 1 First-Stage Sequential Design for Sampling Counties (SNR-Guided)

Require: Current posterior means $\boldsymbol{\mu}_t(\mathbf{x}_c)$ and observation noise variance $\boldsymbol{\Sigma}_t(\mathbf{x}_c)$
Compute scalarized mean and variance for each county c:
$\mu_t = \mathbf{w}^\top \boldsymbol{\mu}_t(\mathbf{x}_c)$
$\sigma_t = \sqrt{\mathbf{w}^\top \boldsymbol{\Sigma}_t(\mathbf{x}_c)\mathbf{w}}$
Compute $\alpha_{\text{SNR}}(c)$
Select county $c^* = \arg\max\ \alpha_{\text{SNR}}(c)$

Based on this selection, we draw S posterior samples in county c^* and use them to construct credible intervals for treatment effects in county c^*, explained next in Section 3.2.

3.2. Second-Stage Sequential Design for Sampling Treatment Conditions

At each iteration, the first-stage sequential design identifies a county c^* for additional sampling, represented by its feature vector $\mathbf{x}_{c^*}$. However, evaluating all $\ell \times \ell$ treatment combinations of n and b levels for the selected county is computationally inefficient. The second stage of the sequential design helps us direct simulation effort toward the treatment condition with the widest predictive credible interval. Given the selected county in the first stage, the second-stage sequential design chooses the most uncertain treatment condition within that county. To identify the most uncertain treatment condition at county c, we use the GPR posterior. Specifically, instead of using $\boldsymbol{\mu}(\mathbf{x}_{c^*}) = [\mu_0(\mathbf{x}_{c^*})\ \mu_n(\mathbf{x}_{c^*})\ \mu_b(\mathbf{x}_{c^*})]^\top$ to estimate the overdose death rate for a specific treatment condition (n, b) in county c^* in equation (1), we draw S posterior samples $\left\{ f_0^{(s)}(\mathbf{x}_{c^*}),\ f_n^{(s)}(\mathbf{x}_{c^*}),\ f_b^{(s)}(\mathbf{x}_{c^*}) \right\}_{s=1}^S$, and use each sample as coefficients in the response function to generate posterior samples for the treatment effect estimates at each treatment condition (n, b) in county c^*:

$$\zeta^{(s)}(n, b \mid c^*) = f_0^{(s)}(\mathbf{x}_{c^*}) + f_n^{(s)}(\mathbf{x}_{c^*}) \cdot n + f_b^{(s)}(\mathbf{x}_{c^*}) \cdot b. \tag{3}$$

Here, we use the S values $\left\{\zeta^{(s)}(n, b \mid c^*)\right\}_{s=1}^S$ as posterior predictive samples for $z(n, b|c^*)$ in equation (1) to evaluate predictive uncertainty. To that end, we compute empirical 95% credible intervals:

$$\text{CI}_{95}(n, b \mid c^*) = \left[\text{Quantile}_{2.5\%}\{\zeta^{(s)}\}, \text{Quantile}_{97.5\%}\{\zeta^{(s)}\}\right], \tag{4}$$

with interval width

$$\text{width}(\text{CI}_{95}(n, b \mid c^*)) = \text{Quantile}_{97.5\%}\{\zeta^{(s)}\} - \text{Quantile}_{2.5\%}\{\zeta^{(s)}\}, \tag{5}$$

and choose the treatment condition with the widest credible interval (Algorithm 2):

$$(n^*, b^*) = \arg\max_{(n,b)} \{\text{width}(\text{CI}_{95}(n, b \mid c^*))\}.$$

Algorithm 2 Second-Stage Sequential Design for Sampling Treatment Conditions

Require: feature vector $\mathbf{x}_{c^*}$ for the county c^* selected by Algorithm 1
  Draw S posterior samples $[f_0^{(s)}(\mathbf{x}_c^*), f_n^{(s)}(\mathbf{x}_c^*), f_b^{(s)}(\mathbf{x}_c^*)]^\top$
  for each treatment condition (n, b) **do**
    Compute predicted outcomes $\zeta^{(s)}(n, b \mid c^*)$ according to equation (3)
    Estimate 95% credible interval $\text{CI}_{95}(n, b \mid c^*)$ using equation (4)
    Compute $\text{width}(\text{CI}_{95}(n, b \mid c^*))$ using equation (5)
  end for
  Select condition $(n^*, b^*) = \arg\max_{(n,b)} \text{width}(\text{CI}_{95}(n, b \mid c^*))$

Together, Algorithms 1 and 2 constitute our two-stage sequential design framework that is grounded in GPR posterior sampling and credible-interval selection and enables efficient allocation of simulations to the most informative counties and treatment conditions to obtain a superior metamodel fit under tight computational constraints.

3.3. Bi-level Metamodel Workflow and Simulation Output Integration

During initialization, we allocate relatively more simulation replications to the baseline treatment condition $(n, b) = (0, 0)$ than to other conditions within the initial batch. This condition anchors baseline overdose mortality and contributes disproportionately to early prediction error, so improved estimation at this point stabilizes subsequent learning of treatment effects. After each simulation batch, the resulting outputs are summarized at the county level by fitting a linear regression to estimate the response-function coefficients. These regression-based coefficient estimates are then incorporated into the contextual-level GPR, which is implemented using the `BoTorch` framework (Balandat et al., 2020). After each simulation batch, the resulting outputs are summarized at the county level by fitting a linear regression to estimate the response-function coefficients. These regression-based coefficient estimates are then incorporated into the contextual-level GPR, which is implemented using the `BoTorch` framework (Balandat et al., 2020).

At the end of each iteration, the GPR posterior is updated with the new regression-based coefficient estimates (as explained in Section 2.2). Each element of $\boldsymbol{\mu}(\mathbf{x}_c)$ corresponds to the posterior mean of a response-function coefficient for the county characterized by feature vector $\mathbf{x}_c$. In the second level, these coefficients are plugged into the response function to get predicted overdose death rates for any treatment condition. Sequential design proceeds iteratively, with the first stage selecting the next county using a signal-to-noise ratio acquisition rule, evaluated from the GPR posterior mean and variance. For that specific county c^*, the second stage uses posterior samples from the GPR to form credible intervals over all treatment conditions, then selects the single condition with the widest interval to run additional simulation samples. The simulation outcomes from these runs are then used to fit a linear regression in Equation 2, yielding updated coefficients for the baseline, naloxone, and buprenorphine effects. These regression coefficients are then appended to the training dataset to update the

GPR posterior using the same `BoTorch` implementation. This loop continues until the simulation sample budget is exhausted or the estimates stabilize. A complete summary of notations used in this workflow is provided in Table 1.

Algorithm 3 Bi-Level Metamodel Workflow

Initialize GPR prior with spatial–socio-economic kernel
Initialize with a small set of county–treatment simulations
repeat
county selection: Evaluate the SNR acquisition function over candidate counties, select the county c* using Algorithm 1
posterior sampling: For county c^* with feature vector $\mathbf{x}_{c^*}$, draw S posterior samples $\{ f_0^{(s)}(\mathbf{x}_{c^*}),\ f_n^{(s)}(\mathbf{x}_{c^*}),\ f_b^{(s)}(\mathbf{x}_{c^*}) \}_{s=1}^{S}$ from the GPR
treatment selection: For each treatment pair (n, b), compute the empirical 95% credible interval; select (n^*, b^*) with the widest interval using Algorithm 2
Simulation and augmentation: Run the simulation at (c^*, n^*, b^*) for a number of replications and append the new observations to the training set
Linear regression: Fit $r(n^*, b^* \mid c^*) = \beta_{0,c^*} + \beta_{n^*,c^*}\, n + \beta_{b^*,c^*}\, b$ using all observed runs for county c^*, and obtain the coefficient vector $\boldsymbol{\beta}_{c^*} = [\beta_{0,c^*}, \beta_{n^*,c^*}, \beta_{b^*,c^*}]$
model update: update the `BoTorch` GPR using $\boldsymbol{\beta}_{c^*}$ as the observation
until the sample budget is reached or the relative error stabilizes

Table 1.: **Summary of notation used in the paper.** By convention, bold lowercase letters (e.g., $\mathbf{x}$) denote vectors, bold uppercase letters (e.g., $\mathbf{K}$) denote matrices, and Greek letters (e.g., μ, Σ) denote GPR parameters.

Symbol	Description
J	Number of interventions (factors)
ℓ	Number of dispensing rate levels per intervention
t	Iteration index in the sequential design algorithm
(n, b)	Treatment condition defined by naloxone level n and buprenorphine level b
$\mathbf{x}_c \in \mathbb{R}^d$	Feature vector for county c (dimension d)
$\beta_0, \beta_n, \beta_b$	Linear regression coefficients
$r(n, b \mid c)$	Linear regression function $\beta_0 + \beta_n n + \beta_b b$ for county c
$z(n, b \mid c)$	Response function $\mu_0 + \mu_n n + \mu_b b$ for county c
$\zeta^{(s)}(n, b \mid c)$	Posterior samples of $z(n, b \mid c)$ (indexed by s)
$\boldsymbol{\Sigma}(\mathbf{x}_c) = \mathrm{diag}[\sigma_0^2(\mathbf{x}_c), \sigma_n^2(\mathbf{x}_c), \sigma_b^2(\mathbf{x}_c)]^\top$	Diagonal matrix of observation noise variances for county c in the heteroscedastic GPR
$\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c), \mu_n(\mathbf{x}_c), \mu_b(\mathbf{x}_c)]^\top$	Response function coefficients modeled by GPR posterior means

4. Implementation and Numerical Results

Our outcome of interest is the cumulative number of overdose deaths over a five-year horizon for each county and treatment condition. We calibrated our OUD model to reproduce county-level opioid mortality patterns over a five-year pre-pandemic study

period 2015-2019, for which we had county-level outcomes and covariates data available. Six counties, Allegheny, Philadelphia, Dauphin, Erie, Columbia, and Clearfield, were calibrated using incremental mixture importance sampling (Menzies et al., 2017). These counties were then used as prototypes to generalize model parameters across the remaining Pennsylvania counties using a feature-based matching procedure.

The matching transfers calibrated model parameters governing outcome progression and behavioral dynamics to counties with similar historical overdose mortality and dispensing-rate trajectories, while county-specific naloxone and buprenorphine dispensing rates are preserved and applied independently as local inputs. This approach avoids the infeasibility of full calibration for all 67 counties while retaining county-level heterogeneity in treatment exposure. Further details of the FRED simulation platform and the OUD model structure are provided in Appendix C, calibration details are given in Appendix D, and the computational infrastructure and implementation details are described in Appendix E.

To rigorously test the metamodel framework, we conducted an exhaustive numerical experiment consisting of 25 treatment conditions (five naloxone levels $\times$ five buprenorphine levels) for each of the 67 counties. Each condition was replicated 1000 times to average out stochastic variation in the agent-based simulation, resulting in more than 1.6 million simulation runs in total. This large baseline experiment serves two purposes: first, to establish a benchmark for comparing the proposed metamodel against brute-force simulation; and second, to highlight that further scaling of the design space is computationally prohibitive without surrogate modeling. Training the full bi-level metamodel, including sequential design and GPR fitting, required approximately two hours of wall-clock time on a Google Colab environment equipped with 8 vCPUs (AMD EPYC, $\sim$2.25 GHz), and 12 GB of RAM, highlighting the practical computational efficiency of the proposed approach relative to exhaustive simulation.

Model accuracy is evaluated using relative error, defined for each county and treatment condition as the absolute difference between the response function prediction and the average held-out simulation output, divided by that average, and then averaged across all counties and treatment conditions. The test set consists of 20% of the simulation replications per county-treatment condition, held out before any model training. Since each replication is an independent draw from the stochastic simulator under a fixed condition, the replications are exchangeable, and the holdout requires no ordering or stratification.

4.1. Sampling Efficiency and Learning Curves

We organize the subsequent empirical evaluation around three interrelated questions that govern the performance of the proposed bi-level metamodel. First, how many simulation runs are required to achieve reliable predictive accuracy under a sequential design? Second, how much complexity is needed in the response function to capture treatment effects without sacrificing interpretability? Third, how should the kernel be designed to support this model complexity, capturing spatial and socio-economic heterogeneity across counties while avoiding overfitting and unstable posterior behavior? The results that follow examine these questions in turn and demonstrate how sample allocation, response-function structure, and kernel design must be jointly balanced to achieve accuracy, efficiency, and robustness.

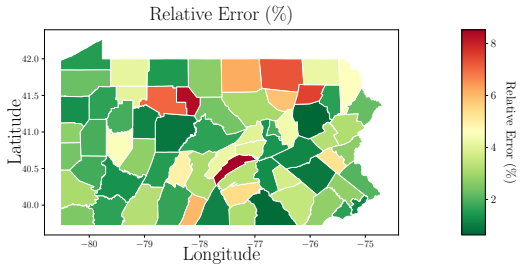
The cumulative allocation of simulation runs across counties is highly uneven, reflecting the adaptive behavior of the sequential design. As shown in Figure 3a, more

than two-thirds of counties require fewer than 150 simulation runs, while a small subset of north-central counties, highlighted in green and yellow, receive substantially more samples, with up to 600 runs per county. These allocations correspond to the state of the model after a total of 10,000 simulation runs.

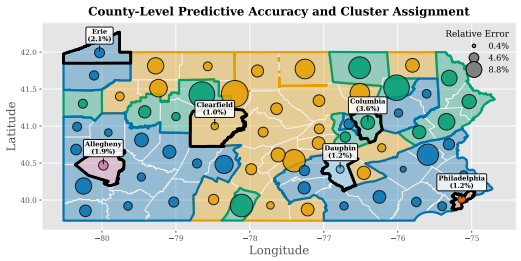
Across most counties, the relative error of overdose mortality predictions is below 5%, indicating high predictive accuracy of the metamodel. Counties with higher relative error coincide with those receiving greater simulation effort, reflecting the adaptive behavior of the sequential design, which concentrates sampling where the model is most difficult to learn until uncertainty is reduced to an acceptable level. This spatial error pattern is shown in Figure 2a, with the corresponding sample allocation illustrated in Figure 3a. Figure 2b further shows that prediction error varies across clusters, with the Clearfield and Columbia clusters exhibiting higher mean relative error, and Figure 2(c)–(d) confirm that errors remain largely uniform across treatment conditions and decrease with county population size.

As a further check of model performance, we examine prediction error across two additional comparisons: calibrated versus non-calibrated counties, and sequentially sampled versus never-selected counties. Figure 2e shows that higher errors within the Columbia and Clearfield clusters are concentrated in small-population counties (Cameron, Wyoming, Sullivan), where simulation variance is inherently higher due to small population size. Figure 2f demonstrates that the five never-selected counties achieve lower mean relative error (1.8% vs. 3.2%), indicating that the SNR acquisition function correctly identified them as low-uncertainty regions and that the GPR generalizes accurately without direct simulation runs.

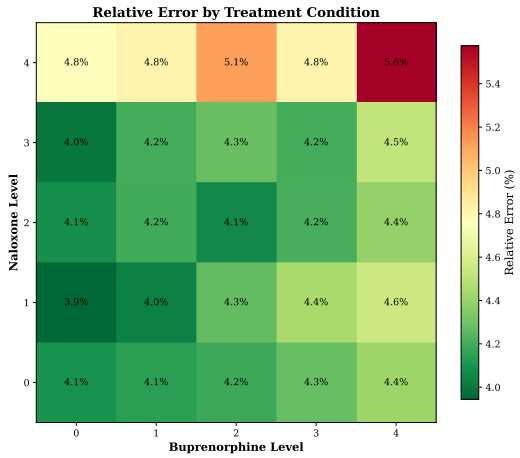
Compared with the approximately 1.6 million simulations required for exhaustive enumeration, the bi-level metamodel framework, combining Gaussian process regression with outcome-level response-function modeling, achieves comparable statewide accuracy using only a fraction of the total simulation budget, demonstrating its suitability for large-scale policy analysis under tight computational constraints.



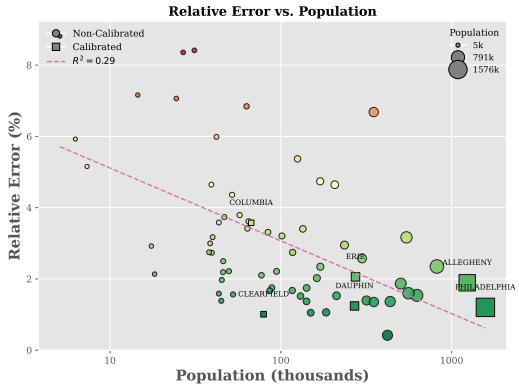
(a) County-level predictive accuracy



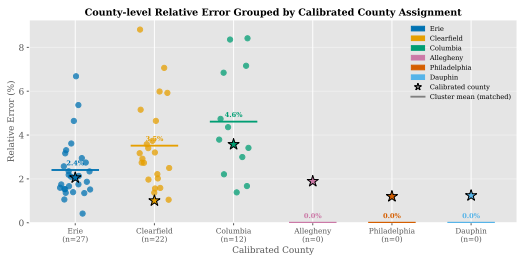
(b) County-level relative error and cluster assignment



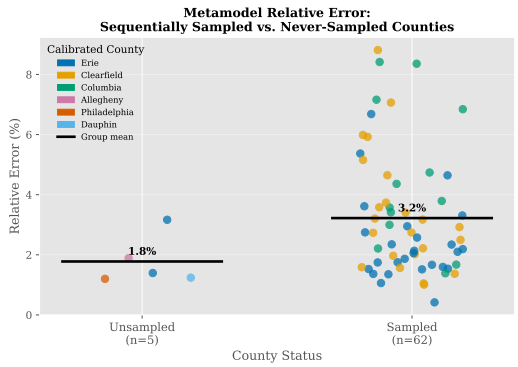
(c) Relative error by treatment condition



(d) Relative error vs. county population size

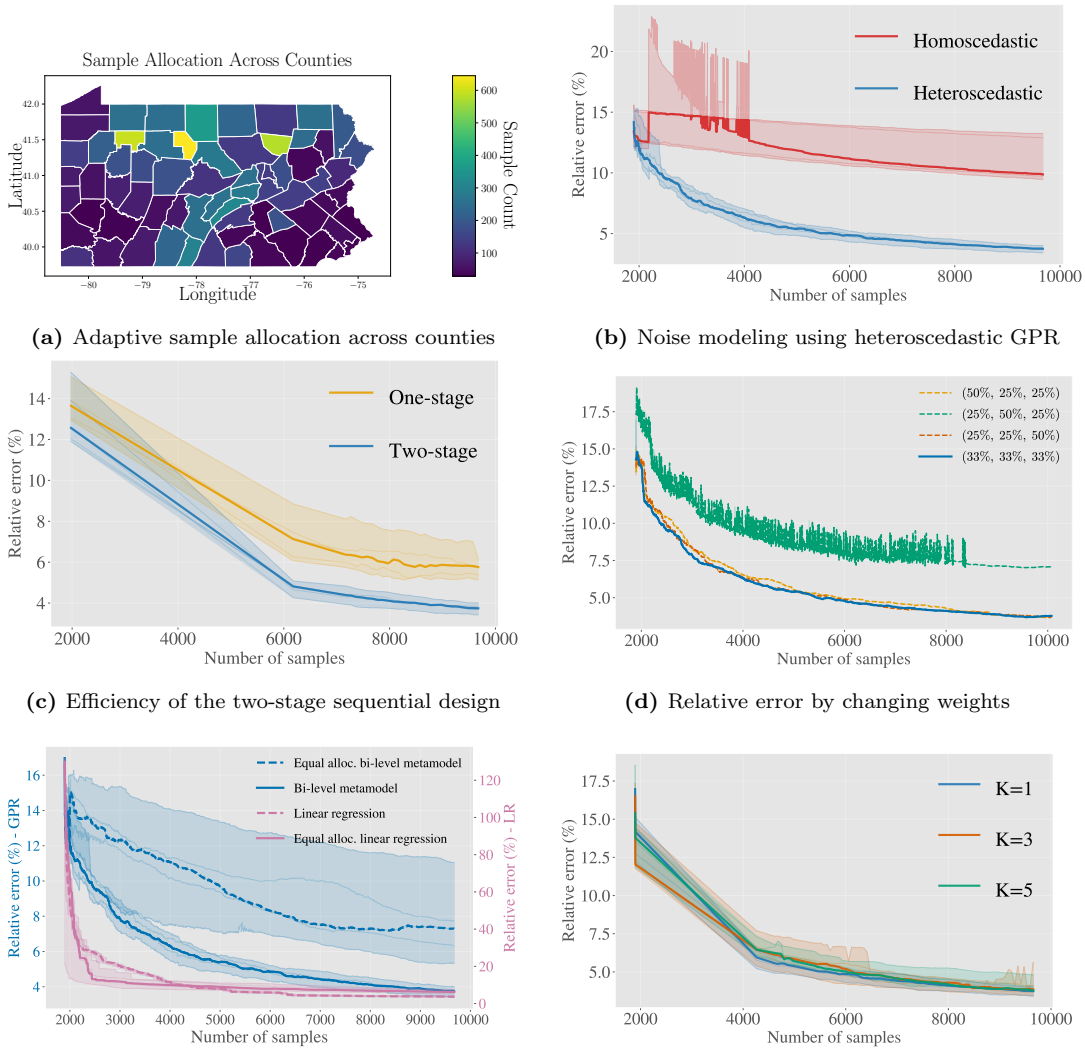


(e) Relative error grouped by calibration assignments



(f) Relative error for sampled vs. unsampled counties

Figure 2: Empirical evaluation of the heteroscedastic noise modeling and sequential design strategies for improving the sample efficiency of the proposed bi-level modeling framework.
Panel (a) shows the county-level predictive accuracy of the metamodel after 10,000 total samples.
Panel (b) shows county-level relative error and cluster assignment, where circle size reflects the magnitude of relative error and fill color indicates the assigned cluster. Counties are grouped into six clusters, each represented by a calibrated county (Allegheny, Philadelphia, Dauphin, Erie, Columbia, and Clearfield) shown with bold black borders. Each county is shaded with the color of its assigned calibrated county.
Panel (c) shows the mean relative error across the 5×5 intervention grid, with slightly higher error at extreme treatment conditions.
Panel (d) shows a negative association between county population and relative error ($R^2 = 0.29$), indicating that the metamodel achieves higher accuracy in more populous counties.
Panel(e) reports county-level relative error grouped by calibrated county assignment.
Panel(f) compares relative error between the 62 sequentially sampled counties and the 5 never-selected counties (Philadelphia, Dauphin, Allegheny, Luzerne, Lancaster).



(a) Adaptive sample allocation across counties

(b) Noise modeling using heteroscedastic GPR

(c) Efficiency of the two-stage sequential design

(d) Relative error by changing weights

(e) Learning curves for different baselines

(f) Effect of batch size on learning curve

Figure 3: Sample efficiency of the bi-level metamodel under alternative noise specifications, sequential design strategies, baselines, and batch sizes.
Panel (a) shows the adaptive allocation of simulation runs across counties. The sequential design concentrates effort on counties with higher posterior uncertainty, as reflected by the uneven distribution of sample counts.
Panel (b) shows ignoring county-level heteroscedasticity when specifying observation noise in the GPR model result in a slow, unstable and inefficient learning behavior (the red learning curve).
Panel (c) compares two sampling strategies: (i) a one-stage sequential design, which selects counties adaptively and then exhaustively simulates all treatment conditions within the selected county, and (ii) the proposed two-stage design that additionally selects treatment conditions based on posterior uncertainty.
Panel (d) reports a sensitivity analysis over five scalarization weight configurations $\mathbf{w} = (w_0, w_n, w_b)$ in the SNR acquisition function, varying the emphasis placed on the intercept (μ_0), naloxone (μ_n), and buprenorphine (μ_b) coefficients. default and that the sequential design is not sensitive to this choice.
Panel (e) compares the sequential bi-level metamodel against three baselines under identical simulation budgets: a sequential linear regression baseline, which assigns each sampled county its own OLS estimate but cannot generalize to unsampled counties compared to the spatial interpolation in the bi-level metamodel; an equal allocation bi-level metamodel, which uses the same GPR framework but distributes simulation runs uniformly across counties and treatment conditions rather than adaptively; and an equal allocation linear regression baseline.
Panel (f) compares the learning curves under batch sizes $K = 1$, $K = 3$, and $K = 5$, where K denotes the number of counties selected per sequential design iteration.

Explicitly modeling heterogeneous observation variance leads to substantially more stable and accurate learning behavior in the GPR. Under the heteroscedastic specification, relative error decreases smoothly as additional samples are incorporated, reflecting consistent posterior updating as uncertainty contracts in counties with increasing numbers of simulation replicates. In contrast, the homoscedastic formulation exhibits noticeably less stable learning, with oscillatory error trajectories at the beginning when sample sizes are small. These fluctuations arise from the constant-variance assumption, under which early or sparsely sampled observations can exert disproportionate influence on the posterior, resulting in mischaracterized uncertainty and irregular updates. Consequently, the homoscedastic model converges more slowly and displays greater variability across sample sets. This contrast is illustrated in Figure 3b, where the limited overlap between the shaded min–max bands further indicates that the advantage of the heteroscedastic model is robust rather than driven by a small number of favorable runs. Overall, these results demonstrate that linking observation variance to sample size and regression uncertainty improves both estimation accuracy and learning stability in sequential simulation settings.

Having characterized how the specification of the observation noise influence learning behavior in Figure 3b, we next examine how the sequential design procedure impacts the rate at which the metamodel improves. Figure 3c compares two variants of the metamodel for a 25-condition array: one that employs the variance-oriented sequential design for selecting the next treatment condition (blue curve) and a baseline that samples all treatment conditions in selected county (orange curve). The two-stage sequential strategy achieves a noticeably steeper decline in relative error during the early stages of training and maintains a consistently lower error across the full range of simulated samples. By concentrating additional runs on the single most informative intervention level within each county, the procedure accelerates convergence and reduces the total number of simulations required to reach a given accuracy threshold. This confirms that adaptive allocation of treatment conditions complements county-level sampling, yielding further gains in overall sample efficiency.

Building on the benefits of the two-stage sequential design, we next examine the sensitivity of the framework to three modeling choices: the scalarization weights in the acquisition function, the contribution of spatial interpolation, and the effect of batch size on learning performance. Figure 3d reports a sensitivity analysis over seven scalarization weight configurations $\mathbf{w}$ in the SNR acquisition function. All configurations converge to comparable relative error by 10,000 samples, confirming that equal weighting is a robust default, and the framework is not sensitive to this choice. The exception is the configuration that places higher weight on the buprenorphine coefficient ($\mathbf{w} = (25\%, 25\%, 50\%)$), which exhibits higher and more unstable relative error throughout training. This is consistent with buprenorphine having a smaller effect size than the intercept and naloxone coefficients, so over-weighting it directs sampling toward counties that are uncertain in a low-magnitude coefficient rather than in the overall outcome.

To benchmark the bi-level metamodel, Figure 3e shows that the sequential bi-level metamodel achieves substantially lower relative error than all three baselines throughout the learning curve, demonstrating that efficiency gains stem from both spatial interpolation across counties and the adaptive sample allocation of the sequential design. Figure 3f examines whether selecting $K > 1$ counties per iteration improves performance. Increasing the batch size to $K = 3$ or $K = 5$ does not reduce relative error relative to the single-county update $K = 1$, and experiments with larger batch sizes confirm the same conclusion. This confirms that one-at-a-time posterior updating

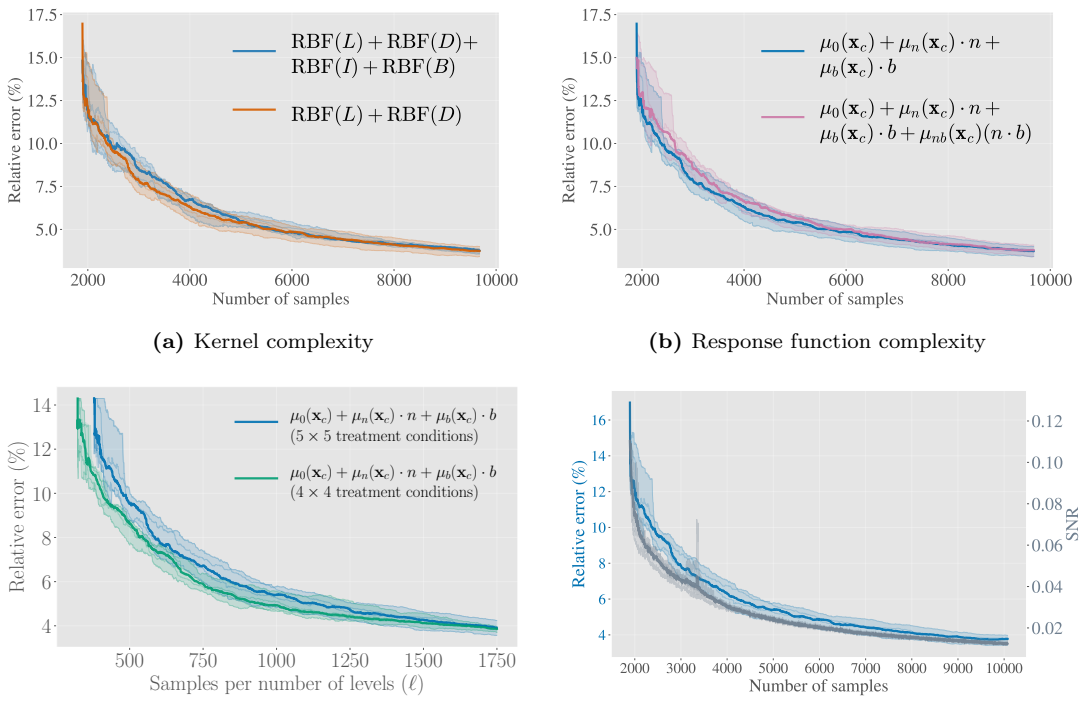
is the most sample-efficient strategy. This is consistent with the sequential nature of the design: each selection benefits from the full posterior update induced by all previous observations, an advantage that is diminished when multiple counties are selected simultaneously on the same posterior surface.

Having established the benefits of modeling heteroscedastic uncertainty and the two-stage sequential design in Figure 2, we next examine the role of model complexity in shaping metamodel performance, i.e., sample efficiency and accuracy. In the following paragraphs, we discuss the effect of GPR kernel and the response-function complexity, as well as the size of the intervention grid on learning performance, shown in Figure 3.

Response-function complexity. To investigate the effect of response function complexity and the trade-offs between sample efficiency and representational capacity, we compared the performance of the main-effects response function in Equation 1 and an interaction-augmented specification within the metamodel framework. The more complex interaction-augmented response function model is specified as follows (compare with the main-effects-only model in Equation (1)):

$$z^{(\text{int})}(n, b|c) = \mu_0(\mathbf{x}_c) + \mu_n(\mathbf{x}_c)n + \mu_b(\mathbf{x}_c)b + \mu_{nb}(\mathbf{x}_c)(n \cdot b). \tag{6}$$

Figure 4b shows the learning curves of two response-function specifications evaluated under the same kernel design, $k(\cdot, \cdot) = \text{RBF}(L) + \text{RBF}(D) + \text{RBF}(I) + \text{RBF}(B)$, where L, D, I, and B denote location (L), population density (D), income (I), and black population percentage (B; see Equation (B2) in Appendix B). Across all sample sizes, the main-effects model attains substantially lower median relative error and exhibits a much narrower performance range. In contrast, the interaction specification shows persistently higher error and greater variability, indicating that the additional interaction term increases model complexity without improving predictive accuracy at the available sample sizes. This behavior reflects a well-known principle: when data are limited, a simpler response function can provide more stable and reliable estimates than more complex alternatives. In our setting, the main-effects model offers the best trade-off between interpretability, precision, and sample efficiency, while the interaction specification would require a larger simulation budget to be estimated reliably.



(a) Kernel complexity

(b) Response function complexity

(c) Design space complexity

(d) SNR trajectory during sequential sampling

Figure 4: Effect of different types of model complexity on sample efficiency of the learning curves. The baseline model shown in blue remains the same across the four panels. **Panel (a)** Learning is slower for the more complex Kernel that combines more features: the blue baseline uses location (L), population density (D), median household income (I), and percent back population (B) versus only L and D used in the orange.
Panel (b) Learning is slower for the more complex response function that models the interaction between the two interventions (in red), compared to the blue baseline that only includes the main effects.
Panel (c) Learning is slower when modeling outcomes in a larger intervention grid (5×5 in blue vs 4×4 in green); however the sample complexity scales with number of levels ℓ for each intervention rather than the grid size ℓ^2, consistent with Ahmed et al., 2024, Theorem 1.
Panel (d) shows the outcome SNR, averaged across all counties and treatment conditions, decaying in parallel with the relative error throughout the sequential sampling process.

Kernel complexity. We next examined how the choice of kernel influences metamodel performance. Figure 4a reports relative-error learning curves for two kernel specifications applied to the same response function. The simpler specification, RBF(L)+RBF(D), captures spatial and demographic variation using only location (L) and population density (D). The more expressive specification, RBF(L) + RBF(D) + RBF(I) + RBF(B), augments this structure with additional socio-economic features, increasing kernel flexibility. At smaller sample sizes, the higher-dimensional kernel exhibits slightly higher error and greater variability, reflecting the increased variance and hyperparameter uncertainty associated with more complex covariance structures under limited data. As additional samples are collected, this disadvantage diminishes: the richer kernel steadily closes the gap and ultimately achieves marginally lower relative error and smoother convergence than the simpler alternative.

Design space and intervention grid size. To assess outcome-level learning behavior under changes in the intervention design space, we evaluated the performance of the metamodel across treatment grids of different sizes. Figure 4c plots the learning

curves of the two–level metamodel for two intervention grids: a 4×4 array (16 treatment conditions, green curve) and a 5×5 array (blue curve). Despite the 56 % increase in design points, the larger grid achieves approximately the same relative error after a comparable number of sequential design samples. Both curves fall rapidly during the first ∼200 simulation runs (per treatment condition) and level off near a relative error of 4% by 350 simulations (per treatment condition), indicating that the county-condition sequential design successfully targets high-uncertainty regions regardless of grid size. These results demonstrate that the metamodel maintains predictive accuracy and sample efficiency even as the intervention design space grows.

Figure 4d overlays the relative error with the outcome SNR, computed as the average signal-to-noise ratio of the response function predictions across all counties and treatment conditions at each sampling iteration. In early iterations, when few counties have been sampled, SNR measures are high, indicating that the GPR posterior remains uncertain relative to the predicted treatment effects. As the sequential design preferentially allocates simulation runs to counties with the highest acquisition SNR, posterior uncertainty contracts rapidly.

The outcome SNR tracks the relative error decline closely throughout the sampling process, falling steeply during the same early phase and gradually stabilizing as predictions converge. This parallel behavior highlights an important operational advantage of the GPR framework: unlike black-box surrogates, the GPR posterior provides well-characterized uncertainty estimates that serve as a reliable proxy for prediction accuracy, enabling system designers to monitor learning curve progress and anticipate convergence without access to ground truth simulation outcomes. The declining outcome SNR serves as an online stopping criterion, allowing decision-makers to determine when sufficient simulation effort has been allocated.

4.2. Localized Treatment Effects

We observed substantial heterogeneity in both baseline overdose mortality and intervention outcome estimates across Pennsylvania counties. Baseline levels differ considerably across regions, as evidenced by large variation in the intercept term μ_0, while the estimated treatment effects for naloxone and buprenorphine are predominantly negative yet vary considerably in magnitude across counties. This pattern indicates that increases in either intervention are generally associated with reductions in overdose deaths, but that the strength of these associations is highly county dependent. In addition, uncertainty in the estimated effects is uneven across regions, with wider credible intervals in counties with more limited simulation or calibration data, reflecting differential information content across the spatial domain. Collectively, these findings suggest that uniform statewide intervention policies are unlikely to be efficient and instead motivate the need for county-specific strategies that account for local baseline mortality and heterogeneous treatment responsiveness, as shown in Figure 5.

To better understand the spatially localized and heterogeneous treatment effects identified above, in Table 2 we report posterior means and 95% credible intervals for the response-function coefficients in the six calibrated counties that we used as prototypes in our modeling framework. Philadelphia exhibits the highest baseline mortality (μ_0), per 100,000 people, but also the strongest naloxone effect, whereas smaller counties such as Columbia and Clearfield show lower baseline and more modest treatment effects.

Robustness to outcome-level model specification. To test the stability of the

Table 2.: Posterior mean and 95% credible intervals for response-function coefficients across selected calibrated counties (predicting overdose deaths per 100,000 people over five years).

County	μ_0 (95% CI)	μ_n (95% CI)	μ_b (95% CI)
Allegheny	88.96 [88.36, 89.56]	−4.26 [−4.48, −4.04]	−5.65 [−5.87, −5.43]
Philadelphia	338.71 [335.68, 341.74]	−25.85 [−26.14, −25.56]	−3.91 [−4.22, −3.60]
Dauphin	216.17 [213.20, 219.14]	−2.09 [−2.38, −1.79]	−9.65 [−9.93, −9.37]
Erie	109.58 [107.33, 111.84]	−3.41 [−3.66, −3.15]	−2.48 [−2.73, −2.23]
Columbia	34.34 [33.43, 35.25]	−0.96 [−1.14, −0.77]	−2.22 [−2.40, −2.03]
Clearfield	22.28 [21.67, 22.89]	−0.93 [−1.02, −0.85]	−0.80 [−0.90, −0.70]

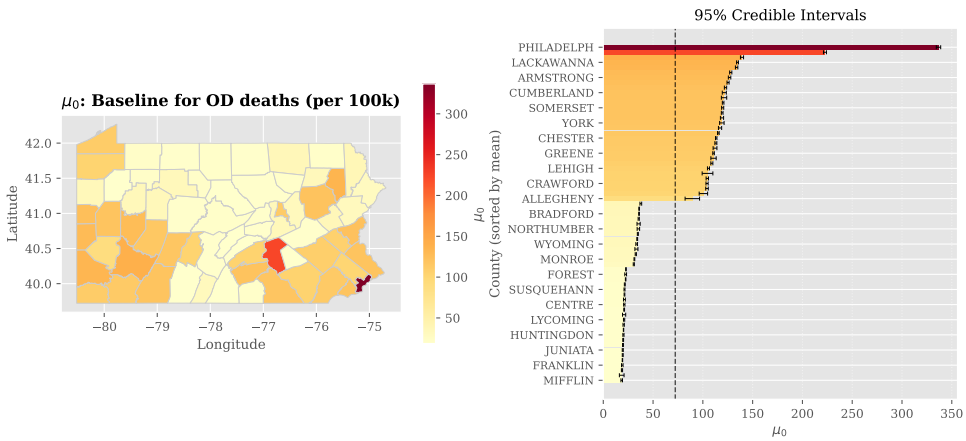
county-level estimates, we compared the naloxone and buprenorphine effect sizes obtained from the bi-level framework under different response function specifications in Equation (6), with and without the interaction term. The magnitude of the estimates obtained from the main-effects-only model were on average larger, but the differences between the estimates from the two model specifications were consistently small in all counties: the mean difference (interaction minus main effects) in the estimates for μ_n was -0.13 (SD $= 0.61$) and for μ_b was -0.25 (SD $= 0.73$). In Figure 6a, we report the differences between the main effect estimates in the two models (with and without the interaction term) with credible intervals for the top 14 counties with the largest-magnitude differences. The majority of intervals span zero, confirming that the main-effects estimates obtained from the combined GPR and response function modeling framework are robust to the specification of an interaction term in the response function.

Figure 6b shows the county-level estimates of the interaction coefficient (μ_{nb}). The interaction effects are small in magnitude with mean $= 0.12$ (SD $= 0.24$, 10th–90th percentile: $[-0.06, 0.45]$) compared to the main effects for naloxone and buprenorphine, whose means are -2.86 (SD $= 3.40$, 10th–90th percentile: $[-5.50, -0.62]$) and -2.80 (SD $= 2.89$, 10th–90th percentile: $[-7.43, -0.37]$), respectively. Most credible intervals for estimated interactions include zero, supporting the main-effects specification used in the primary analysis, and consistent with our observations of parallel trends in the factorial analysis in Figure B1.

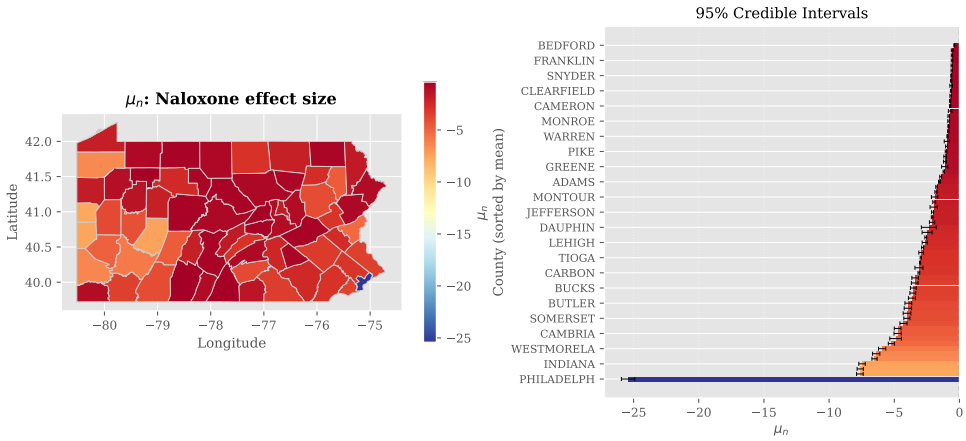
To further validate the independent-output GPR formulation, we examined the pairwise correlations among iterative coefficient updates $\Delta\beta_m = \beta_m^{(t)} - \beta_m^{(t-1)}$ across all sequential design iterations. The results show that $\mathrm{corr}(\Delta\beta_n, \Delta\beta_b) = -0.04$, confirming that the two treatment effect coefficients update nearly independently throughout training. The stronger correlations observed between the intercept and treatment coefficients, $\mathrm{corr}(\Delta\beta_0, \Delta\beta_n) = -0.63$ and $\mathrm{corr}(\Delta\beta_0, \Delta\beta_b) = -0.58$, reflect the well-known intercept-slope trade-off in OLS regression.

From localized treatment effects to locally-tailored policies. Our results show that uniform, state-wide intervention strategies are unlikely to be effective, as baseline overdose mortality and treatment effect magnitudes vary substantially across counties, necessitating locally tailored policies. Recent simulation-based studies of opioid interventions find that the combined effects of harm-reduction strategies (i.e., naloxone distribution), medications for opioid use disorder (i.e., buprenorphine treatment), and diversion-based policies depend on local conditions, and that these interventions can exhibit synergistic effects when deployed jointly, such that the most effective policy combinations vary across communities (Cerdá et al., 2024; White & Albert, 2025). This conclusion is consistent with work in operations and policy analytics

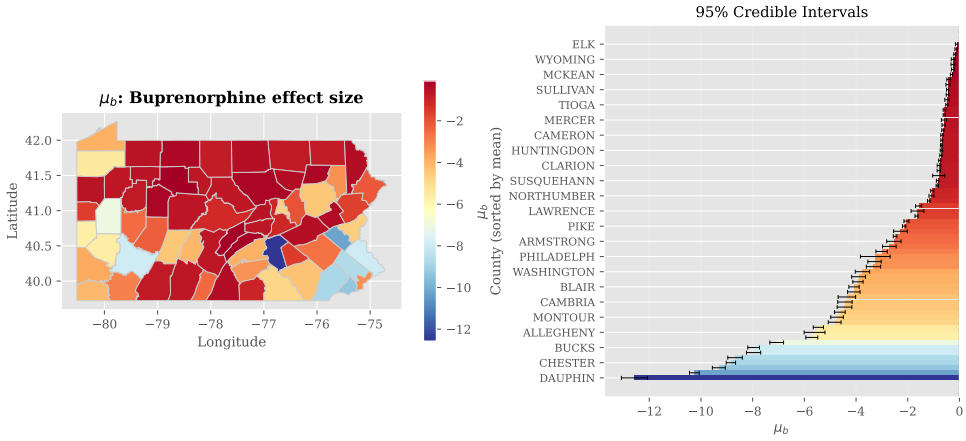
demonstrating that policies optimized at an aggregate level can perform poorly when applied uniformly across heterogeneous regions, whereas explicitly accounting for spatial and contextual heterogeneity improves resource allocation efficiency and policy performance (Luo & Stellato, 2024). Complementary epidemiological evidence further shows that opioid overdose risk and intervention effectiveness vary across geographic areas due to differences in local drug environments, population characteristics, and healthcare access, and that population-averaged analyses may not accurately reflect these local patterns (Dodson et al., 2018). Together, these findings motivate policy evaluation frameworks that support locally adaptive strategies informed by heterogeneous baseline mortality and treatment effects.



(a) Estimated baseline overdose mortality (μ_0) per 100,000 people

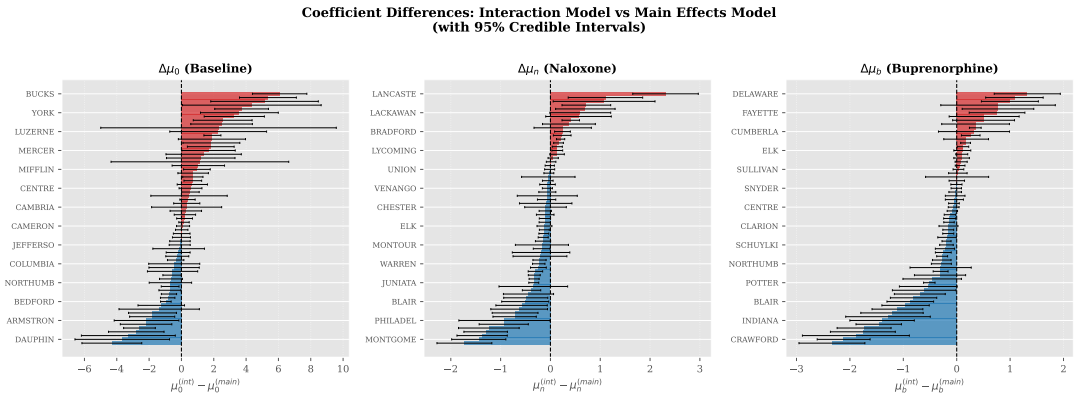


(b) County-specific estimates of naloxone effect sizes (μ_n)

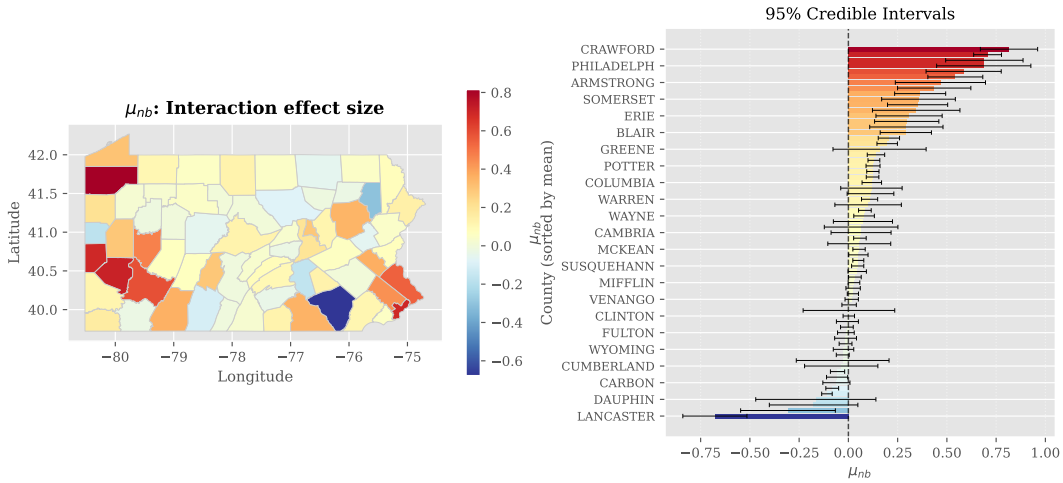


(c) County-specific estimates of buprenorphine effect sizes (μ_b)

Figure 5: Posterior summaries of the GPR-estimated response-function coefficients across Pennsylvania counties. Each panel shows the posterior mean (left) and the corresponding 95% credible intervals sorted by posterior mean (right). The three coefficients represent: (a) μ_0: intercept, (b) μ_n, the change in overdose deaths per 100,000 people associated with a one-level (25%) increase in naloxone dispensing relative to the county's baseline naloxone dispensing rate; and (c) μ_b, the corresponding change associated with a one-level (25%) increase in buprenorphine dispensing relative to the county's baseline buprenorphine dispensing rate. Together, these coefficients enable the estimation of overdose deaths per 100,000 people for any specified combination of naloxone and buprenorphine treatment levels in all counties.



(a) Coefficient differences between the interaction model and the main-effects model, computed as $\Delta\mu = \hat{\mu}^{(\text{int})} - \hat{\mu}^{(\text{main})}$, with 95% credible intervals for baseline mortality $\Delta\mu_0$ on the left, naloxone effect $\Delta\mu_n$ in the middle, and buprenorphine effect $\Delta\mu_b$ on the right. Counties are sorted by the magnitude of the differences and only the top 14 are shown.



(b) Posterior estimates of the interaction coefficient μ_{nb} across Pennsylvania counties, showing the spatial distribution of posterior means on the left and 95% credible intervals on the right, shown only for the 23 largest magnitudes.

Figure 6: Robustness analysis comparing estimates obtained for the main-effects model in Equation (1) **and the interaction-augmented response function in Equation** (6)**.** Panel (a) confirms that adding the interaction term does not substantially alter the main effect estimates, as the majority of coefficient differences have credible intervals spanning zero. Panel (b) shows that the estimated interaction effects are small in magnitude.

Computational complexity and scalability. The $O(N^3)$ cost of GPR posterior inference, arising from the inversion of the $N \times N$ kernel matrix $\mathbf{K}$ (Rasmussen & Williams, 2006, Chapter 2), applies with respect to N, the number of counties incorporated into the training set, rather than the number of simulation replicates. In our setting, $N = 67$ Pennsylvania counties across all iterations: the acquisition function selects among the same fixed set of locations at every step, so the kernel matrix retains its 67×67 dimensions and the $O(N^3)$ cost remains a constant per-iteration overhead. Scaling to all U.S. counties ($N \approx 3{,}143$) would increase the per-iteration inference cost substantially as it grows as $O(n^3)$. In this regime, sparse GP approximations (Titsias, 2009; Katzfuss & Guinness, 2021) would be necessary, replacing the full kernel matrix with a smaller set of $M \ll N$ inducing points (auxiliary locations chosen to summarize

the information in the full training set) and reducing the inference cost to $O(M^2N)$.

Separately, larger county populations require more simulation replicates to achieve the same variance reduction in the regression coefficients, increasing the simulation cost independently of the GP. Regarding kernel structure, the composite RBF kernel encoding spatial proximity and socioeconomic similarity is expected to remain applicable across U.S. counties since the same county-level features (location, income, population density, racial composition) are available nationwide. However, extending to multiple states would require careful consideration of regional heterogeneity, such as state-level policy environments and drug supply patterns, which may necessitate additional kernel components to capture between-state variation not explained by the current feature set.

5. Conclusions

This study presents a novel bi-level metamodeling framework to address the computational challenges of evaluating multi-level intervention policies across a high-dimensional, spatially heterogeneous domain. The methodology integrates a Gaussian process regression (GPR) surrogate with a response function to efficiently emulate a complex opioid use disorder (OUD) simulation model outcome. The contextual-level GPR, equipped with a custom composite kernel that incorporates geographic and socio-economic features, learns the simulation outcome by capturing spatial correlations and county-level heterogeneity across communities. The outcome-level response function then provided an interpretable and computationally efficient mechanism for estimating outcomes under any intervention combination within the evaluated grid.

The bi-level metamodel, when coupled with proposed two-stage sequential design, achieves high predictive accuracy and sampling efficiency. Across all counties, the metamodel achieves relative errors of approximately 5% or less while requiring fewer than 2% of the simulation runs required for exhaustive enumeration of the county–intervention space. The GPR component enables efficient learning of spatial structure and cross-county heterogeneity in the response-function coefficients, while the outcome-level model translates these learned relationships into accurate predictions across the full intervention grid. As a result, the framework not only reproduces simulation outcomes with high fidelity but also recovers meaningful county-specific differences in baseline overdose mortality and intervention effect sizes for naloxone and buprenorphine. These empirical findings confirm that the two-stage sequential design effectively concentrates computational effort where it is most informative, allowing the metamodel to scale to large, high-dimensional policy spaces without sacrificing interpretability or accuracy.

Beyond methodological advances, the broader impact of this framework lies in its potential to inform policy decisions at the county and state levels. We envision that the proposed framework could serve as the computational backbone of an interactive decision-support tool designed to support policy exploration workflows used by public health agencies, such as the Centers for Disease Control and Prevention (CDC). Such a tool would enable decision makers to explore projected outcomes under alternative treatment allocations across multiple counties and years without requiring exhaustive simulation. By leveraging the metamodel, county-level projections of overdose mortality and OUD prevalence could be generated for arbitrary intervention combinations, along with associated credible intervals to support uncertainty-aware comparisons across strategies.

Future Work

In the current framework, the GPR provides point estimates for the coefficients of the outcome-level response function that links naloxone and buprenorphine levels to overdose deaths. A promising research direction involves transitioning these coefficients not as fixed values but as random variables with full probability distributions. In such a formulation, the GPR posterior would define informative prior distributions over the response-function coefficients, which could then be updated using additional simulation or observational data within a Bayesian model. The Bayesian model would then combine these priors with observed simulation data through Bayes' theorem to obtain posterior distributions for each coefficient. This hierarchical formulation propagates uncertainty from the GPR into the final coefficient estimates, supports sequential updating as additional data are collected, and provides a principled basis for model assessment. Overall, the framework transforms the metamodel from a predictive surrogate into a probabilistic inference framework that yields comprehensive uncertainty quantification for policy evaluation.

Another fruitful avenue for future inquiry lies in enriching the contextual GPR with a correlated multi-output structure. In the present implementation, each response-function coefficient is modeled using an independent Gaussian process, a choice that simplifies the implementation while remaining well-justified given that the three coefficients are always combined jointly at the outcome level through the response function. In future work, the GPR framework could be extended to incorporate cross-output covariance through coregionalization or related multi-output GPR constructions. Such structures would allow the GPR to learn explicit dependencies among the latent functions and may further improve sample efficiency and predictive performance in regions where treatment effects exhibit shared spatial or socio-economic structure.

Another important direction for future research is to extend the current spatially and socio-economically aware framework into a full spatio–temporal setting. The present metamodel is based on cross-sectional data, treating counties as static units, characterized by geographic centroids and socio-economic features over the five-year study period. In reality, epidemic trajectories are shaped not only by the underlying disease dynamics but also by policy shifts and emerging drug trends which evolve over time. Subsequently, intervention outcomes can be better optimized when modeled over time and allowed to adapt to the changing dynamics. Incorporating the time dimension would transform the model into a spatio-temporal GPR, where each county's coefficient vector depends not only on spatial–socio-economic features but also on time.

A natural starting point would be to extend the GPR kernel to jointly model spatial and temporal variation, capturing smooth evolution of the response-function coefficients over time while preserving the efficiency of the current framework. The heteroscedastic noise model and two-stage sequential design would extend naturally to this setting, with the first stage selecting the most uncertain county-year combination at each iteration rather than a county alone, enabling the metamodel to capture how treatment effectiveness evolved as the opioid epidemic changed over the 2015–2019 period and to propagate temporal uncertainty into county-level policy recommendations.

Data and Code Availability

Our analysis relies on multiple data streams that collectively capture OUD dynamics at the county level. Monthly county-level dispensing rates for prescription opioids, nalox-

one, and buprenorphine were obtained from the IQVIA dataset. These data reflect prescriptions dispensed across retail, mail-order, and long-term care pharmacies. This data is not publicly available. Access can be requested through IQVIA at `https://www.iqvia.com/insights/the-iqvia-institute/available-iqvia-data`. County-level overdose death counts were collected from the CDC Wide-Ranging Online Data for Epidemiologic Research, identified using International Classification of Diseases, 10th Revision (ICD–10) codes for opioid-related poisoning (X40–X44, X60–X64, X85, Y10–Y14, T40.0–T40.4, T40.6). In addition, fentanyl seizure rates were obtained from the National Forensic Laboratory Information System (NFLIS) and can be accessed at `https://www.nflis.deadiversion.usdoj.gov/`. All code used to implement the GPR metamodel and generate the figures presented in this paper is publicly available at `https://github.com/abdulrahmanfci/gpr-metamodel`.

Acknowledgment

This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided, and by the National Science Foundation under grant agreement CMMI-2240408. Specifically, this work used the HTC and VIZ clusters, which are supported by NIH award number S10OD028483. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation. We thank Philippe Giabbanelli and Hamdi Kavak for helpful feedback.

Disclosure statement

The authors report no conflict of interest.

Consent and Approval Statement

This study has been exempt from the requirement for approval by University of Pittsburgh review board.

References

Ahmed, A. A., Rahimian, M. A., & Roberts, M. S. (2023a). Estimating treatment effects using costly simulation samples from a population-scale model of opioid use disorder. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 1–4).

Ahmed, A. A., Rahimian, M. A., & Roberts, M. S. (2023b). Inferring epidemic dynamics using gaussian process emulation of agent-based simulations. In *Proceedings of Winter Simulation Conference (WSC)*.

Ahmed, A. A., Rahimian, M. A., & Roberts, M. S. (2024). Optimized model selection for estimating treatment effects from costly simulations of the us opioid epidemic. In *2024 annual modeling and simulation conference (annsim)* (pp. 1–13).

Ahmed, S., Sarfraz, Z., & Sarfraz, A. (2022). *A changing epidemic and the rise of opioid-stimulant co-use* (Vol. 13). Frontiers Media SA.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., & Bakshy, E. (2020). BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in neural information processing systems 33*. Retrieved from `http://arxiv.org/abs/1910.06403`

Ball, F., & House, T. (2017). Heterogeneous network epidemics: real-time growth, variance and extinction of infection. *Journal of Mathematical Biology*, *75*(3), 577–619.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825–848.

Bhatt, S., Cameron, E., Flaxman, S. R., Weiss, D. J., Smith, D. L., & Gething, P. W. (2017). Improved prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal of The Royal Society Interface*, *14*(134), 20170520.

Blanco, C., Wiley, T. R., Lloyd, J. J., Lopez, M. F., & Volkow, N. D. (2020). America's opioid crisis: the need for an integrated public health approach. *Translational psychiatry*, *10*(1), 167.

Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Buckingham-Jeffery, E., Isham, V., & House, T. (2018). Gaussian process approximations for fast inference from infectious disease data. *Mathematical biosciences*, *301*, 111–120.

CDC. (2024). *Understanding drug overdose and deaths.* Retrieved from `https://www.cdc.gov/drugoverdose/epidemic/index.html` ([Accessed: 2024-08-10])

Centers for Disease Control and Prevention. (2025). *Drug overdose mortality: Stats of the states.* Retrieved from `https://www.cdc.gov/nchs/state-stats/deaths/drug-overdose.html` (Accessed: March 26, 2026)

Cerdá, M., Hamilton, A. D., Hyder, A., Rutherford, C., Bobashev, G., Epstein, J. M., ... others (2024). Simulating the simultaneous impact of medication for opioid use disorder and naloxone on opioid overdose death in eight new york counties. *Epidemiology*, *35*(3), 418–429.

Cerdá, M., Jalali, M. S., Hamilton, A. D., DiGennaro, C., Hyder, A., Santaella-Tenorio, J., ... Keyes, K. M. (2021). A systematic review of simulation models to track and address the opioid crisis. *Epidemiologic reviews*, *43*(1), 147–165.

Cheng, H., McGovern, M. P., Garneau, H. C., Hurley, B., Fisher, T., Copeland, M., & Almirall, D. (2022). Expanding access to medications for opioid use disorder in primary care clinics: an evaluation of common implementation strategies and outcomes. *Implementation Science Communications*, *3*(1), 72.

Cheng, L.-F., Dumitrascu, B., Darnell, G., Chivers, C., Draugelis, M., Li, K., & Engelhardt, B. E. (2020). Sparse multi-output gaussian processes for online medical time series prediction. *BMC medical informatics and decision making*, *20*(1), 152.

Ciccarone, D. (2021). The epidemiology of the opioid overdose epidemic in the united states. In *The opioid epidemic and infectious diseases* (pp. 1–10). Elsevier.

Conti, S., & O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of statistical planning and inference*, *140*(3), 640–651.

Dodson, Z. M., Enki Yoo, E.-H., Martin-Gill, C., & Roth, R. (2018). Spatial methods to enhance public health surveillance and resource deployment in the opioid epidemic. *American journal of public health*, *108*(9), 1191–1196.

D'Onofrio, G., O'Connor, P. G., Pantalon, M. V., Chawarski, M. C., Busch, S. H., Owens, P. H., ... Fiellin, D. A. (2015). Emergency department–initiated buprenorphine/naloxone treatment for opioid dependence: a randomized clinical trial. *Jama*, *313*(16), 1636–1644.

Fearnhead, P., Giagos, V., & Sherlock, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics*, *70*(2), 457–466.

Fisher, A., Adhikari, B., Zhai, C., Morgan, J. E., Mago, V. K., & Giabbanelli, P. J. (2020). Predicting the resource needs and outcomes of computationally intensive biological simulations. In *2020 spring simulation conference (springsim)* (pp. 1–12).

Forrester, A., Sobester, A., & Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Frazier, P. I. (2018). Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems* (pp. 255–278). Informs.

Garnett, M. F., & Miniño, A. M. (2024). *Drug overdose deaths in the united states, 2003–2023*.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., ... others (2013). Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health*, *13*(1), 1–14.

Guclu, H., Kumar, S., Galloway, D., Krauland, M., Sood, R., Bocour, A., ... Potter, M. (2016). An agent-based model for addressing the impact of a disaster on access to primary care services. *Disaster medicine and public health preparedness*, *10*(3), 386–393.

Gutierrez, J.-J. G., Lau, E., Dharmapalan, S., Parker, M., Chen, Y., Álvarez, M. A., & Wang, D. (2024). Multi-output prediction of dose–response curves enables drug repositioning and biomarker discovery. *npj Precision Oncology*, *8*(1), 209.

Irvine, M. A., Oller, D., Boggis, J., Bishop, B., Coombs, D., Wheeler, E., ... others (2022). Estimating naloxone need in the usa across fentanyl, heroin, and prescription opioid epidemics: a modelling study. *The Lancet Public Health*, *7*(3), e210–e218.

Jalal, H., Buchanich, J. M., Roberts, M. S., Balmert, L. C., Zhang, K., & Burke, D. S. (2018). Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016. *Science*, *361*(6408), eaau1184.

Jenkins, R. A. (2021). The fourth wave of the us opioid epidemic and its implications for the rural us: a federal perspective. *Preventive medicine*, *152*, 106541.

Katzfuss, M., & Guinness, J. (2021). A general framework for vecchia approximations of gaussian processes. *Statistical Science*, *36*(1), 124–141.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Langmüller, A. M., Chandrasekher, K. A., Haller, B. C., Champer, S. E., Murdock, C. C., & Messer, P. W. (2024). Gaussian process emulation for exploring complex infectious disease models. *medRxiv*, 2024–11.

Li, C.-Y., Rakitsch, B., & Zimmer, C. (2022). Safe active learning for multi-output gaussian processes. In *International conference on artificial intelligence and statistics* (pp. 4512–4551).

Liebschutz, J. M., Crooks, D., Herman, D., Anderson, B., Tsui, J., Meshesha, L. Z., ... Stein, M. (2014). Buprenorphine treatment for hospitalized, opioid-dependent patients: a randomized clinical trial. *JAMA internal medicine*, *174*(8), 1369–1376.

Lim, T. Y., Stringfellow, E. J., Stafford, C. A., DiGennaro, C., Homer, J. B., Wakeland, W., ... others (2022). Modeling the evolution of the us opioid crisis for national policy development. *Proceedings of the National Academy of Sciences*, *119*(23), e2115714119.

Luo, J., & Stellato, B. (2024). Frontiers in operations: Equitable data-driven facility location and resource allocation to fight the opioid epidemic. *Manufacturing & Service Operations Management*, *26*(4), 1229–1244.

Marotta, P. L., Hunt, T., Gilbert, L., Wu, E., Goddard-Eckrich, D., & El-Bassel, N. (2019). Assessing spatial relationships between prescription drugs, race, and overdose in new york state from 2013 to 2015. *Journal of psychoactive drugs*, *51*(4), 360–370.

Menzies, N. A., Soeteman, D. I., Pandya, A., & Kim, J. J. (2017). Bayesian methods for calibrating health policy models: a tutorial. *Pharmacoeconomics*, *35*(6), 613–624.

Moraga, P. (2023). *Spatial statistics for data science: theory and practice with r*. Chapman and Hall/CRC.

Raftery, A. E., & Bao, L. (2010). Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. *Biometrics*, *66*(4), 1162–1173.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning.* MIT press.

Reiker, T., Golumbeanu, M., Shattock, A., Burgert, L., Smith, T. A., Filippi, S., ... Penny, M. A. (2021). Emulator-based bayesian optimization for efficient multi-objective calibration of an individual-based model of malaria. *Nature communications*, *12*(1), 7212.

Sanchez, S. M., Sanchez, P. J., & Wan, H. (2020). Work smarter, not harder: A tutorial on designing and conducting simulation experiments. In *2020 winter simulation conference (wsc)* (pp. 1128–1142).

Sawe, S. J., Mugo, R., Wilson-Barthes, M., Osetinsky, B., Chrysanthopoulou, S. A., Yego, F., ... Galárraga, O. (2024). Gaussian process emulation to improve efficiency of computationally intensive multidisease models: a practical tutorial with adaptable r code. *BMC Medical Research Methodology*, *24*(1), 26.

Scheidell, J. D., Townsend, T. N., Zhou, Q., Manandhar-Sasaki, P., Rodriguez-Santana, R., Jenkins, M., ... others (2024). Reducing overdose deaths among persons with opioid use disorder in connecticut. *Harm reduction journal*, *21*(1), 103.

Senanayake, R., O'Callaghan, S., & Ramos, F. (2016). Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 30).

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics* (pp. 567–574).

Volkow, N. D., & Blanco, C. (2021). The changing opioid crisis: development, challenges and opportunities. *Molecular psychiatry*, *26*(1), 218–233.

Wheaton, W. (2012). *Us synthetic population database 2005–2009: Quick start guide. rti international.*

White, V. M., & Albert, L. A. (2025). Evaluating diversion and treatment policies for opioid use disorder. *IISE Transactions on Healthcare Systems Engineering*, 1–28.

Wilson, J., Hutter, F., & Deisenroth, M. (2018). Maximizing acquisition functions for bayesian optimization. *Advances in neural information processing systems*, *31*.

Zang, X., Bessey, S. E., Krieger, M. S., Hallowell, B. D., Koziol, J. A., Nolen, S., ... others (2022). Comparing projected fatal overdose outcomes and costs of strategies to expand community-based distribution of naloxone in rhode island. *JAMA network open*, *5*(11), e2241174.

Zheng, T., Keyes, K., Ji, S., Calderon, A., Wu, E., Doogan, N. J., ... others (2025). Opioid use disorder prevalence in 57 new york counties from 2017 to 2019: A bayesian evidence synthesis. *Drug and alcohol dependence*, *267*, 112548.

Zimmer, C., & Yaesoubi, R. (2020a). Influenza forecasting framework based on gaussian processes. In *International conference on machine learning* (pp. 11671–11679).

Zimmer, C., & Yaesoubi, R. (2020b). Influenza forecasting framework based on gaussian processes. In *International conference on machine learning* (pp. 11671–11679).

Zimmer, C., Yaesoubi, R., & Cohen, T. (2017). A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLoS computational biology*, *13*(1), e1005257.

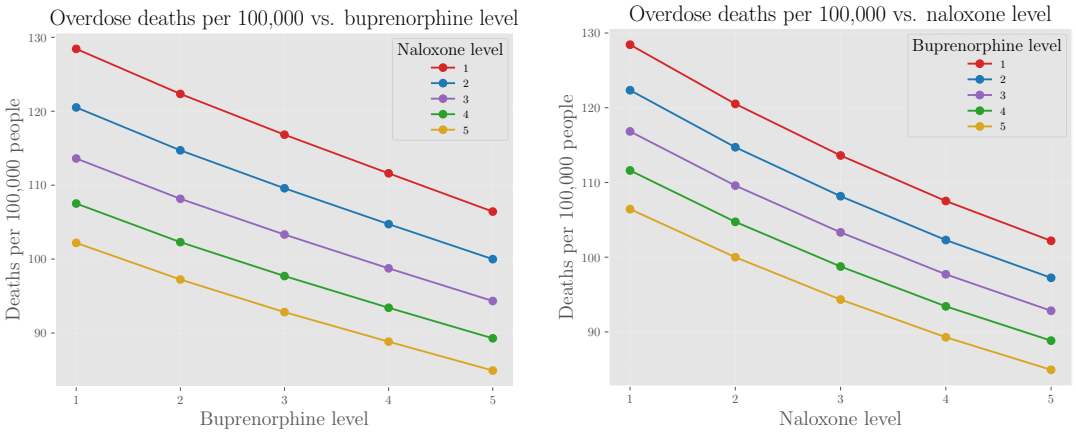
Appendix A. Other Related Work

Estimating a single treatment effect from a population-scale simulation is feasible. However, the challenge intensifies when multiple treatment effects must be evaluated, particularly across a wide design space. In such settings, the computational cost of running repeated simulations becomes a significant barrier. Metamodels offer a promising solution by approximating simulation outputs without requiring full evaluations at every point. Among various approaches, Gaussian Process regression (GPR) has emerged as a widely used metamodeling technique due to its flexibility, ability to quantify uncertainty, and strong performance in interpolating between sparse observations.

Several studies have demonstrated the utility of GPR in modeling epidemic dynamics and public health processes. For instance, Fearnhead et al. (2014) use GPR models to infer reaction rates in biological networks, showing how their method generalizes to a broad class of reaction systems relevant to epidemic modeling. Senanayake et al. (2016) develop a spatiotemporal GPR framework to forecast influenza trends using large-scale data, incorporating multiple kernels to capture seasonality, non-stationarity, and long- and short-term patterns. Zimmer and Yaesoubi (2020a) apply GPR regression to influenza forecasting using CDC data, benchmarking its performance against alternative methods. Similarly, Zimmer et al. (2017) employ GPR models to calibrate epidemic parameters such as the duration of infectiousness and expected future cases in real time. Ball and House (2017) estimate the mean and variance of infections in an SIR network model using a GPR informed by a branching process covariance structure. Buckingham-Jeffery et al. (2018) propose a GPR-based framework for SIR and SEIR models, comparing multiple GPR variants for parameter estimation.

Appendix B. Model Selection in the Two-Level Framework

Response function design. To assess the interaction between naloxone and buprenorphine, we use two factorial plots showing the cumulative deaths per 100,00 people in Pennsylvania (statewide) over the five-year study period (2015-2019) for different treatment levels (Figure B1): panel (a) shows mean overdose deaths per 100,000 versus buprenorphine level with separate lines for naloxone levels 1-5; panel (b) reverses the roles. In both panels, the lines are nearly parallel with no systematic crossings, indicating additive main effects and little evidence of interaction. This supports our choice of the simple main-effects model in Equation 1 for our main analysis. This observation is consistent with the small magnitudes of the estimated interaction coefficients μ_{nb} in our robustness checks (Figure 6b).



(a) Overdose deaths per 100,000 people in Pennsylvania vs. buprenorphine level after five years; separate lines show naloxone levels 1–5. Lines are nearly parallel (no crossings), indicating no interaction.

(b) Overdose deaths per 100,000 people in Pennsylvania vs. naloxone level after five years; separate lines show buprenorphine levels 1–5. Trends are again near-parallel, indicating no interaction.

Figure B1: Factorial plots over the 5×5 intervention grid. The near-parallel trends in both figures provide no visual evidence of interaction between naloxone and buprenorphine. A simple main-effects model, $z(n, b \mid c) = \mu_0(\mathbf{x}_c) + \mu_n(\mathbf{x}_c) \cdot n + \mu_b(\mathbf{x}_c) \cdot b$, can therefore be sufficient for outcome-level metamodeling (outcomes shown are cumulative five-year overdose deaths, averaged over 500 simulation replications).

GPR Kernel Design. The kernel function evaluates the covariance structure between any two input feature vectors, denoted by $\mathbf{x}_c$ and $\mathbf{x}_{c'}$ where $\mathbf{x}_c \in \mathbb{R}^d$ represents the d-dimensional vector of county-level covariates used by the Gaussian process. Different kernel functions may be employed depending on the characteristics of the underlying process being modeled.

The radial basis function (RBF) kernel is one of the most widely used covariance functions in Gaussian process modeling and is defined as

$$k_{\text{RBF}}(\mathbf{x}_c, \mathbf{x}_{c'}) = \exp(-\frac{d(\mathbf{x}_c, \mathbf{x}_{c'})^2}{2l^2}), \tag{B1}$$

where l is the length scale parameter and $d(\cdot, \cdot)$ is the Euclidean distance between the input points. RBF kernels, also known as squared exponential kernels, are widely used in GPR modeling due to their ability to model smooth, nonlinear relationships (Rasmussen & Williams, 2006, Chapter 4). In prior work, RBF kernels have been used to encode similarity in continuous variables such as time (Gramacy, 2020), geographic distance (Banerjee et al., 2008), or patient-level covariates in healthcare studies (Brochu et al., 2010).

In constructing our custom kernel, we incorporate four radial basis function (RBF) kernels. Candidate socio-economic features were first identified based on their potential relevance to explaining variation in overdose mortality across counties, including unemployment, poverty, primary care physicians rate, rurality index, and racial composition. We then applied a greedy feature selection procedure, sequentially adding features that reduced out-of-sample prediction error and discarding those that did not yield improvement. The final kernel structure includes four RBF components, each corresponding to a selected contextual feature: the county's geographic location (via

centroid coordinates), median household income, population density, and the percentage of black residents, as follows:

$$\begin{aligned} k(\mathbf{x}_c, \mathbf{x}_{c'}) =& k_{RBF}(\mathbf{x}_{c,1:2}, \mathbf{x}_{c',1:2}) + k_{RBF}(\mathbf{x}_{c,3}, \mathbf{x}_{c',3}) \\ &+ k_{RBF}(\mathbf{x}_{c,4}, \mathbf{x}_{c',4}) + k_{RBF}(\mathbf{x}_{c,5}, \mathbf{x}_{c',5}), \end{aligned} \tag{B2}$$

where $\mathbf{x}_{c,1:2}$ corresponds to the county centroid coordinates (latitude and longitude), $\mathbf{x}_{c,3}$ denotes median household income, $\mathbf{x}_{c,4}$ denotes population density, and $\mathbf{x}_{c,5}$ denotes the percentage of Black residents. Each RBF component has a distinct length-scale parameter (l) that is optimized to maximize marginal likelihood on data, enabling feature-specific adaptation in smoothness and complexity across contextual covariates.

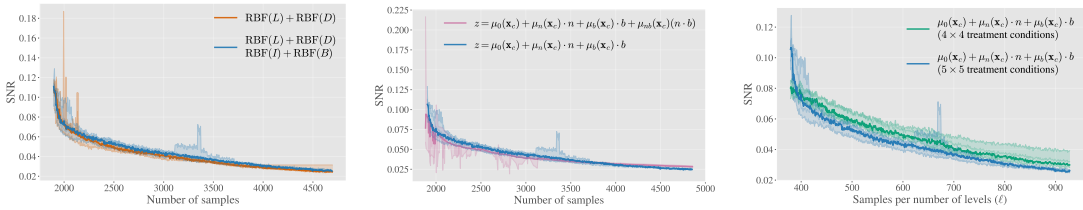
Table B1 reports the relative error and outcome SNR for all kernel configurations considered during greedy feature selection, evaluated at approximately 3,000 simulation samples. The outcome SNR is computed as the average signal-to-noise ratio (σ/μ) of response function predictions drawn from the GP posterior across all counties and treatment conditions at each iteration, and serves as an operationalizable diagnostic for evaluating metamodel uncertainty without requiring held-out ground-truth data. Configurations with substantially elevated outcome SNR, such as the five-feature kernels, indicate numerical instability and are discarded regardless of their relative error.

Table B1.: Outcome SNR and relative error after approximately 3,000 simulation samples. Feature codes: L = location, D = population density, I = income, B = Black population percentage, U = unemployment rate, R = rurality index.

Kernel design	**Features**	**Outcome SNR**	**Rel. Error (%)**
RBF(L)	L	0.0502	9.51
RBF(L) + RBF(I)	L, I	0.0576	9.92
RBF(L) + RBF(D)	L, D	0.0472	8.66
Matérn(L) + RBF(D)	L, D	0.0503	9.59
RBF(L) + RBF(D) + RBF(I) + RBF(B)	L, D, I, B	0.0529	8.97
RBF(L) + RBF(D) + RBF(I) + RBF(U)	L, D, I, U	0.0502	10.06
RBF(L) + RBF(D) + RBF(I) + RBF(B) + RBF(U)	L, D, I, B, U	0.6860	57.69
RBF(L) + RBF(D) + RBF(I) + RBF(B) + RBF(R)	L, D, I, B, R	3.4289	43.02

Outcome SNR as an Alternative Evaluation Metric. Figure B2 replicates the three complexity comparisons of Figure 4, kernel complexity, response function complexity, and design space complexity, using outcome SNR in place of relative error. For kernel complexity, the conclusions are fully consistent: the more complex kernel

exhibits higher outcome SNR, mirroring Figure 4(a). For response function complexity, the interaction model exhibits lower outcome SNR than the main-effects model, because the additional interaction term $\mu_{nb}(\mathbf{x}_c)(n \cdot b)$ provides richer characterization of the treatment response, reducing relative posterior uncertainty. Similarly, for design space complexity, the 5×5 grid exhibits lower outcome SNR than the 4×4 grid, as the larger intervention space provides more treatment conditions from which to average posterior uncertainty, yielding a more precise response function estimate. In both cases, the model with lower outcome SNR also achieves lower relative error, so the metric correctly identifies the better-performing model. Outcome SNR therefore remains a reliable operational metric for online model monitoring and stopping decisions from the GP posterior without requiring held-out data, but should be interpreted in context: lower SNR indicates either higher certainty or richer model structure.



(a) Kernel complexity **(b)** Response function complexity **(c)** Design space complexity

Figure B2: Outcome SNR trajectories replicating the three complexity comparisons of Figure 4, kernel complexity (a), response function complexity (b), and design space complexity (c), confirming that outcome SNR yields consistent conclusions with relative error across all three comparisons.

Appendix C. FRED Agent-Based Modeling of Opioid Use Disorder

The Framework for Reconstructing Epidemiological Dynamics (FRED) is an open-source, agent-based modeling platform developed to simulate the spread of infectious diseases (Grefenstette et al., 2013). While originally designed to enhance understanding of epidemic dynamics, FRED has proven effective as a decision-support tool for public health planning and intervention policy development. A key strength of FRED is its use of synthetic populations derived from real U.S. Census data, which enhances the realism and credibility of its simulations (Guclu et al., 2016). Moreover, FRED builds on the developers' extensive experience with earlier simulation models, allowing it to overcome many limitations found in previous approaches.

Synthetic Population. FRED assigns each individual in the simulation to a specific geographic region, using the U.S. synthetic population database developed by RTI International (Wheaton, 2012), which provides detailed, geographically stratified demographic data. For every agent, FRED generates comprehensive socioeconomic and demographic attributes (e.g., age, education level, income) as well as health-related characteristics (e.g., symptom severity, infection history, vaccination records). Each agent is linked to a specific household and is also assigned to institutions such as schools, workplaces, or prisons. These geographic assignments implicitly encode spatial relationships, including the distance between agents and their assigned locations. Agents in FRED are capable of making individual-level decisions related to health behaviors, such as accepting vaccinations, staying home when ill, or keeping a sick child home from school.

Opioid Use Disorder Model. The rise of Opioid Use Disorder (OUD) has led to a substantial increase in morbidity and mortality in the United States, with over 100,000 overdose deaths (ODDs) reported in 2023 alone. Opioids, including prescription opioids, heroin, and synthetic opioids, are the leading cause of these deaths (CDC, 2024). Jalal et al., 2018 analyze the opioid crisis over a span of more than 40 years and found that the current wave of overdose deaths is part of a long-term trend that has persisted across several decades. These observations highlight the importance of developing a robust model to understand the dynamics of OUD. The Public Health Dynamics Laboratory (PHDL) at the University of Pittsburgh has developed simulation models for several infectious diseases and, through the Centers for Disease Control and Prevention (CDC) funding, created the OUD model used in this study.

The OUD model is based on a set of health states (i.e., stages), where an individual transitions from one state to another based on specific probabilities. These transition probabilities were estimated using literature and by calibrating the model to actual overdose death rates. At the start of the simulation, agents begin in different health states. The agents may transition to other health states. For example, a non-use may move to a prescribed opioid use state (i.e., receiving a prescription from an accredited physician) or to an opioid misuse state. From either of these states, the agent can then transition into the OUD state, which may lead to ODD, another cause of death, or entry into treatment.

To formally describe agent transitions within the OUD simulation model, we use logistic regression to define three key probabilities that govern agent movement across health states. These transition probabilities are linked to intervention or environmental covariates, reflecting policy levers and contextual risks. Each probability is expressed in terms of a logistic regression model.

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 \cdot \text{opioid dispensing rate} \tag{C1}$$

$$\log\left(\frac{p_2}{1-p_2}\right) = \beta_2 + \beta_3 \cdot \text{buprenorphine dispensing rate} \tag{C2}$$

$$\log\left(\frac{p_3}{1-p_3}\right) = \beta_4 + \beta_5 \cdot \text{fentanyl seizure rate} - \beta_6 \cdot \text{naloxone dispensing rate} \tag{C3}$$

Equation (C1) models the transition from the nonuser state to prescription opioid use, with the opioid dispensing rate serving as a key predictor. Equation (C2) defines the likelihood of an individual in the OUD state entering treatment as a function of buprenorphine availability. Finally, Equation (C3) captures the probability of an overdose death, incorporating both fentanyl seizure rate and naloxone dispensing rate to represent the balance between risk and harm reduction. These transition equations serve as the foundation for modeling agent behavior under varying intervention scenarios and contextual risk environments.

The objective is to evaluate the effect of interventions, primarily involving two medications: Naloxone and Buprenorphine. Naloxone is an antidote used to reverse opioid overdose, while Buprenorphine is a treatment medication used to support recovery and help individuals return to the nonuse state. The availability of these medications defines the intervention level, and the number of ODDs in a given geographic area is treated as the intervention outcome.

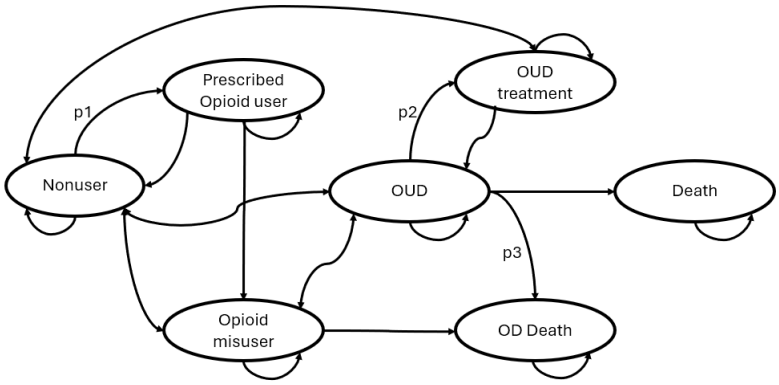


Figure C1: State transition diagram for the OUD model. Transition probabilities p_1, p_2, and p_3 are defined by Equations (C1)-(C3), respectively.

Appendix D. Calibrating the Opioid Use Disorder Agent-Based Model in Different Counties

Model calibration is a critical step in simulation-based studies, particularly when the model includes parameters that are difficult to observe or directly measure. In the context of our OUD simulation model, calibration refers to the process of adjusting unobserved transition probabilities so that the simulated outputs closely match real-world data—most notably, historical overdose death rates. Accurate calibration ensures that simulation results are both credible and reflective of the complex dynamics observed in actual populations.

In this study, we employ Incremental Mixture Importance Sampling (IMIS) (Raftery & Bao, 2010; Menzies et al., 2017) as the calibration algorithm. IMIS is a Bayesian technique that combines the strengths of importance sampling and adaptive proposal distributions. It incrementally builds a mixture of proposal distributions that efficiently explore the high-probability regions of the posterior. This makes IMIS particularly well-suited for models with complex, multimodal likelihood surfaces and moderate-dimensional parameter spaces. The calibration process begins by defining a prior distribution over the uncertain model parameters and specifying a set of target statistics derived from observed data, typically, annual opioid overdose deaths by county and year. The likelihood function measures how well a given parameter configuration reproduces these observed outcomes. IMIS iteratively samples from and updates the proposal distribution to concentrate on regions of the parameter space with high posterior probability.

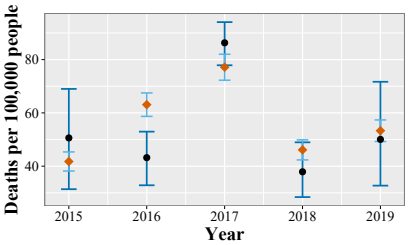
Because calibration is computationally intensive and requires a large number of simulation runs for each parameter set, we limit the calibration to a representative subset of counties. Specifically, we select six counties, Allegheny, Philadelphia, Erie, Dauphin, Clearfield, and Columbia, based on population size and death trends. These are chosen to represent three types of counties: large (Allegheny and Philadelphia), medium (Erie and Dauphin), and small (Clearfield and Columbia). Each pair of counties is selected to reflect varying geographic and epidemiological characteristics. This decision strikes a balance between computational feasibility and the need for generalizable insights across diverse county profiles. To evaluate calibration accuracy, we compared modeled outcomes against observed county-level overdose mortality and examined posterior convergence of the transition-related coefficients (Equations (C1)-(C3)). Figure D1

shows that, across all six calibrated counties, the model closely matches observed mortality patterns and yields well-identified posterior distributions for the calibrated parameters.

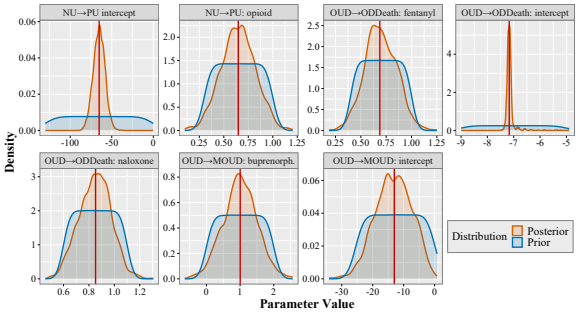
We emphasize that calibration is an expert-driven and resource-intensive process. Each calibration attempt involves multiple simulation replications, validation against mortality curves, visual inspection of trajectories, and iterative refinement of assumptions. Due to these demands, full calibration for all counties is infeasible within reasonable time and resource constraints. Therefore, we select six prototype counties for full calibration, stratified by population size and geography. The calibrated parameters for these counties are then generalized to the remaining counties via a similarity-based assignment procedure. Given the effort required, we perform full calibration only for these six counties and use a similarity-based assignment procedure to generalize the calibrated parameters to non-calibrated counties.

Generalizing calibrated parameters using county Similarity. For each county, we summarize opioid-related trends over 2015–2019 using a small set of features that capture both overall levels and their temporal changes. These features are used to assign each non-calibrated county to a calibrated county with the closest opioid-related covariates. Specifically, we compute summary measures of overdose mortality and treatment dispensing that reflect average magnitude and the estimated slope of each time series over the study period, rather than year-by-year values. Overdose mortality is represented by two quantities: the average overdose death rate over the study period and the estimated slope of overdose mortality over time. Treatment and supply indicators (opioid dispensing, naloxone, buprenorphine, and fentanyl seizures) are summarized by the magnitude of their estimated time-series slopes. The county population is included as an additional contextual feature. These features are standardized across counties and combined into a feature vector for each county. Using this representation, each non-calibrated county is assigned to the most similar calibrated prototype county by minimizing Euclidean distance in the feature space. This matching emphasizes similarity in relative patterns and trends rather than absolute levels.

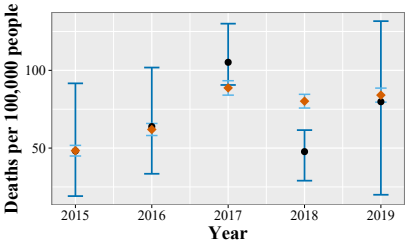
The resulting mapping enables the transfer of calibrated transition-related coefficients from the six prototype counties to all remaining counties, while preserving each county's observed dispensing rates and contextual inputs, thereby eliminating the need for county-by-county calibration and substantially reducing computational and manual effort. Figure D2 summarizes this parameter generalization across Pennsylvania, illustrating how non-calibrated counties inherit transition-related coefficients from their most similar calibrated prototype county.



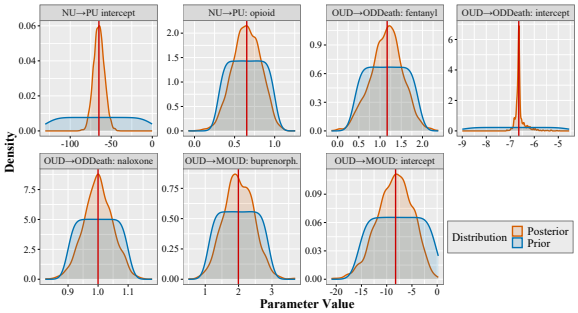
(a) Allegheny calibration fit



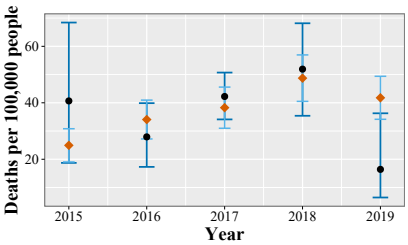
(b) Allegheny posterior distributions of calibrated model coefficients



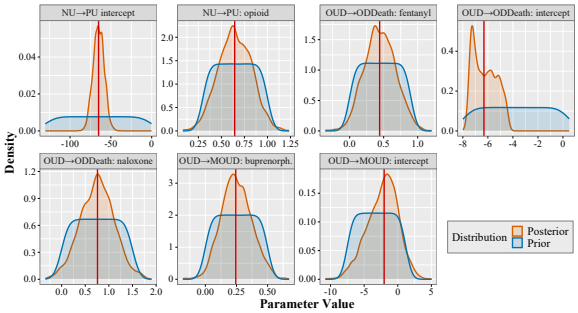
(c) Philadelphia calibration fit



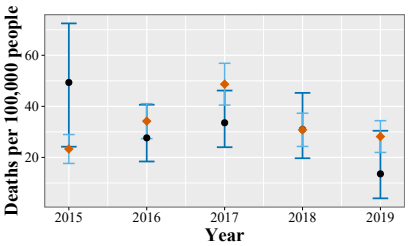
(d) Philadelphia posterior distributions of calibrated model coefficients



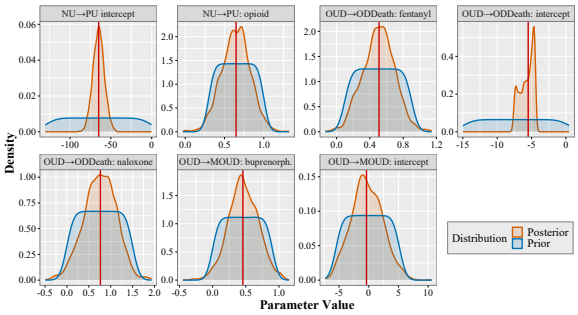
(e) Dauphin calibration fit



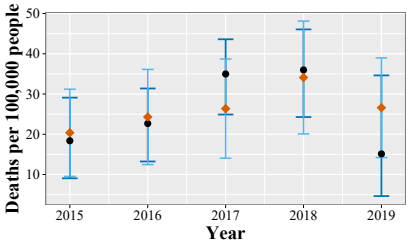
(f) Dauphin posterior distributions of calibrated model coefficients



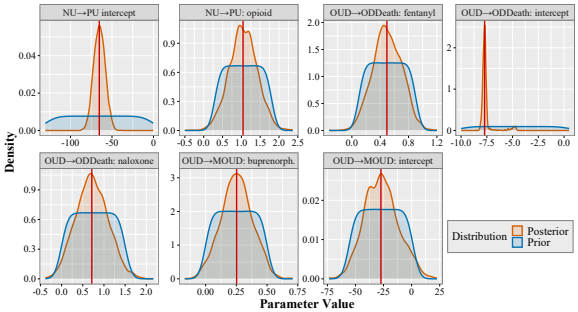
(g) Erie calibration fit



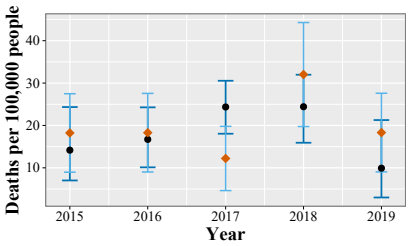
(h) Erie posterior distributions of calibrated model coefficients



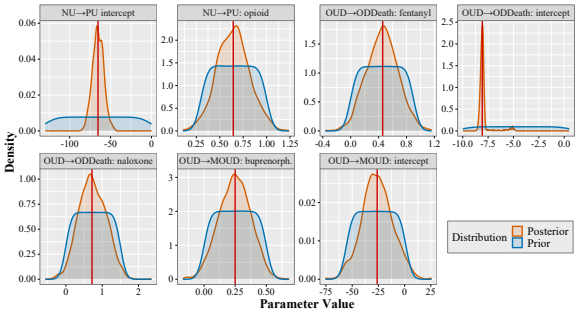
(i) Columbia calibration fit



(j) Columbia posterior distributions of calibrated model coefficients



(k) Clearfield calibration fit



(l) Clearfield posterior distributions of calibrated model coefficients

Figure D1: Calibration performance and parameter identification for the opioid use disorder model. Left panels show model fit to observed county-level overdose mortality targets, where orange points denote observed mortality estimates with associated confidence intervals, and black points with blue credible intervals represent model outputs. While right panels display posterior distributions of the transition-related coefficients defined in Equations (C1)-(C3). across the six calibrated counties (Allegheny, Philadelphia, Dauphin, Erie, Columbia, and Clearfield).

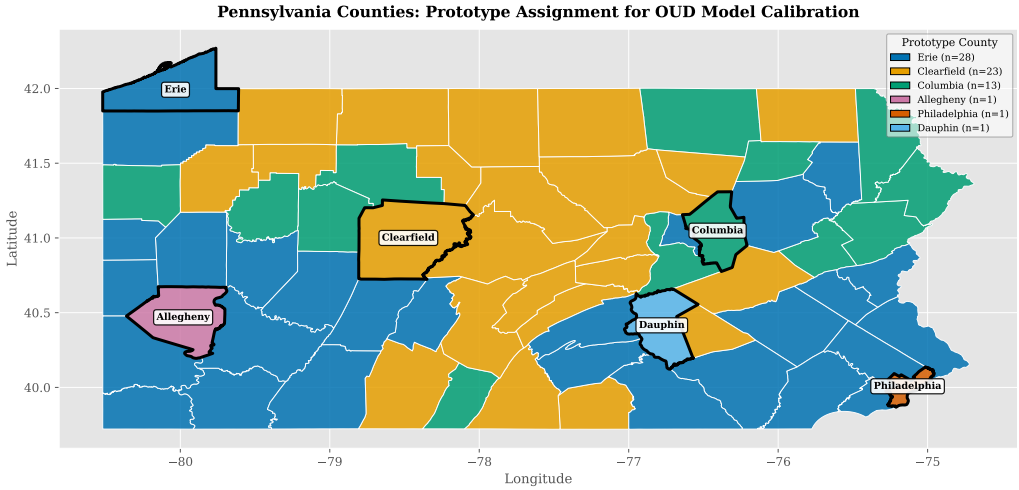


Figure D2: Generalization of calibrated parameters across Pennsylvania. Six counties (highlighted with bold borders) were fully calibrated; remaining non-calibrated counties were matched to the nearest calibrated one based on similarity in overdose mortality patterns and treatment dispensing trends (opioid, naloxone, buprenorphine, and fentanyl seizure rates). Legend shows calibrated prototype county and number of assigned counties (n).

Appendix E. Implementation Details and Computational Constraints

The simulation experiments are executed on the University of Pittsburgh's Center for Research Computing (CRC) cluster, using the Simple Linux Utility for Resource Management (SLURM) job arrays for high-throughput scheduling. Our study involves simulating 67 counties and 25 treatment conditions, each with a target of 1000 replications. This yields more than 1.6 million individual simulation runs. The high replication count serves two purposes: (i) it delivers a low-variance "brute-force" benchmark against which to judge metamodel accuracy, and (ii) it highlights the prohibitive cost of exhaustive simulation. Each simulation utilizes the FRED platform, which models agent-level interactions across synthetic populations with complex disease transmission and treatment dynamics, making each run both memory- and compute-intensive.

The CRC cluster imposes several job submission constraints that introduce bottlenecks in execution. One such constraint is the wall-clock time penalty: jobs requesting longer wall-clock limits are deprioritized in the queue, which significantly delays execution for long-running tasks. Unfortunately, our simulations, especially for densely populated counties or high-treatment levels, require extended runtimes to complete due to the stochastic nature and agent-level detail of the FRED model. As a result, we must balance the specification between a long enough wall-clock to avoid job termination and a short enough request to avoid queuing delays. Another challenge lies in memory management. SLURM enforces strict limits on per-job memory usage, and insufficient allocation can lead to job eviction, segmentation faults, or incomplete logs. On the other hand, excessive memory requests increase wait times and reduce overall cluster utilization. To address this, we use heuristics based on population size, treatment intensity, and historical memory profiles to dynamically scale resource requests. However, occasional resubmission and manual intervention are still required to recover from failures and adjust job configurations. Storage constraints further complicate large-scale simulation. Each simulation generates outputs including agent histories, event logs, and aggregated outcomes. With over a million simulations, the cumulative storage footprint becomes significant. We address this by routinely compressing outputs, writing only summary statistics when feasible, and periodically deleting intermediate files once they are processed into metamodel training datasets.