Computationally Efficient Estimation of Localized Treatment Effects in High-Dimensional Design Spaces using Gaussian Process Regression

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] 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

ARTICLE HISTORY
Compiled December 30, 2025

ABSTRACT
Epidemiological models are widely used to study population-level health dynamics, and population-scale agent-based simulations have proven particularly effective in capturing the complex interactions between individual behaviors and environmental factors. In the context of the opioid epidemic, such models provide a critical tool for evaluating how alternative intervention strategies affect overdose outcomes across heterogeneous communities, supporting the design of locally tailored policies consistent with the principles of precision public health. However, as the number of interventions and their levels increase, the total number of possible combinations of the interventions grows exponentially, becoming prohibitively large for simulation to evaluate each combination exhaustively. In this paper, we develop a two-level metamodel framework, comprising a Gaussian process regression (GPR) followed by a response function, and apply it to efficiently estimate county-level treatment effects on overdose deaths per 100,000 people. At the contextual level, we train a GPR on a sample set of simulation output, learning a smooth mapping from intervention parameters to simulated outcome for each county. To ensure robust performance, we engineer input features, including demographic and socioeconomic covariates, and employ a composite kernel that combines a radial basis function (RBF) component for smooth variation with a periodic component to capture the spatial structure of counties. At the outcome level, the posterior means of the GPR are combined through a response function to generate outcome (i.e., overdose death rate) estimates for a given treatment condition. Using a calibrated opioid epidemic simulation model informed by county-level overdose mortality and treatment dispensing data, our two-level metamodel achieves low relative error ($\sim 5\%$) averaged across all treatment conditions and counties. Compared to exhaustive simulation, this reduced the number of runs by more than 90% while maintaining accuracy in estimating county-level treatment effects. In contrast to classical designs such as fractional factorial or Latin hypercube sampling, our approach leverages spatial correlations and posterior uncertainty to direct sampling more efficiently. This yields a practical decision-support tool for policymakers, enabling rapid evaluation of alternative resource-allocation strategies to mitigate the opioid epidemic in local communities, while also providing a general analytical framework to guide precision public health interventions in broader epidemic settings.

KEYWORDS
metamodel; large-scale simulation; sequential design; model selection;

*Correspondence to: rahimian@pitt.edu

1. Introduction

Simulation modeling has become an essential tool for supporting policy decisions in healthcare, public health, and epidemiology. Such models enable researchers to represent complex population dynamics, incorporate behavioral feedback, and evaluate intervention strategies under realistic conditions. A central challenge in epidemic modeling, however, is heterogeneity: disease dynamics and intervention effects often vary substantially across locations and subpopulations. For example, influenza transmission and vaccine effectiveness differ across regions and age groups (Halloran et al., 2008), measles outbreaks depend strongly on local immunity gaps (Bosetti et al., 2020), and opioid overdose dynamics vary across counties due to differences in drug supply, socioeconomic conditions, and treatment access. As a result, effective policy design requires evaluating intervention strategies that are tailored to specific communities rather than relying on uniform, one-size-fits-all solutions.

This heterogeneity poses a fundamental challenge for simulation-based policy analysis. When interventions operate at multiple levels, and their effects vary locally, analysts must evaluate large combinations of treatment conditions across many geographic units. The computational cost of running such simulations increases rapidly with the size of the input space, especially when spatial structure and multi-component interventions are involved. Factorial or fractional factorial designs have been used in public health simulation studies to explore intervention combinations (Giabbanelli & Crutzen, 2013), but these approaches are practical only when the number of interventions and factor levels is small, because the design size grows exponentially and becomes too large to evaluate in full. Our proposed framework addresses this limitation by enabling efficient exploration of a much larger intervention set, while still capturing interactive effects across counties and treatments.

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 settings that were not directly simulated, typically through interpolation within the sampled design space, 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). Foundational work Kennedy & O'Hagan, 2001 and subsequent extensions by Conti & O'Hagan, 2010 demonstrated how GPR can emulate expensive computer models while propagating parameter uncertainty. In related work, Salle & Yıldızoğlu, 2014 showed that coupling a nearly orthogonal Latin hypercube design with a GPR provides an efficient surrogate for high-dimensional computational models, enabling accurate sensitivity analysis and exploration of parameter spaces with far fewer simulation runs than conventional Monte-Carlo sampling.

To improve learning efficiency, metamodeling is often coupled with sequential design. Rather than selecting all simulation runs in advance, sequential design adaptively chooses the next points to evaluate based on current model predictions. When the metamodel can be trained with relatively few runs, the entire training set may be produced in a single batch without exceeding the simulation budget. However, when more runs are required or when simulations are computationally costly, sequential

design provides a practical strategy by adaptively selecting the most informative next points to evaluate based on current model predictions. This process focuses simulation resources on the most informative areas of the input space. Acquisition functions such as predictive variance, entropy reduction, or expected improvement are used to guide this selection (Frazier, 2018; Wilson et al., 2018; Balandat et al., 2020). These methods are particularly useful when the number of input combinations is large or when the simulation cost is high (Fisher et al., 2020).

Epidemiological models are one of the key application domains for these techniques. Different modeling approaches, including compartmental and agent-based models, have been used to simulate the spread of infectious diseases and evaluate mitigation policies, often at the national or subnational scale. Similar modeling frameworks have been applied to opioid use disorder (OUD) to study the effects of public-health interventions, including naloxone distribution, a harm-reduction strategy that reverses opioid overdoses, and medications for opioid use disorder (MOUD), such as buprenorphine, which mitigate disease burden (Cerdá et al., 2021; Lim et al., 2022). For instance, each intervention is varied across a finite number of dispensing rate levels; if a county dispenses 1,000 naloxone kits per 100,000 people at baseline, a higher level might represent a 10% increase to 1,100 kits per 100,000 people. With five dispensing rate levels for both naloxone and buprenorphine, a single county requires evaluation of 25 treatment conditions. Identifying the policy that yields the greatest improvement in epidemic outcomes requires determining the most effective combination of dispensing rates for the two interventions across counties.

This structure leads to a rapid expansion in the number of design points. If each of J interventions can be assigned ℓ levels, the total number of treatment combinations grows as ℓ^J. When simulations must be run for dozens of counties, the total number of required runs increases multiplicatively; with fifty counties and five interventions at five levels each, a full factorial design entails $50 \times 5^5 = 156{,}250$ base configurations, and this count rises further when multiple stochastic replications are needed. 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 require a trade-off: one can capture higher-order interactions and policy synergies, but only at substantial computational cost; or reduce the number of runs, but at the expense of overlooking some of these effects. Combined with the spatial correlation across counties, the stochastic variability of epidemic simulation outputs, and the exponentially growing input space, this trade-off makes large-scale simulation modeling particularly challenging.

For epidemiological models, simulation output often appears as a vector of coefficients describing how each treatment affects the outcome (e.g., overdose death rates). These coefficients are 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. Recent advances in multi-output and spatial GPR show that shared-kernel constructions can capture such correlations, yet their application in large-scale health simulations has been relatively limited (Alvarez et al., 2012; Binois et al., 2018). Further discussion of related work is provided in Appendix B.

Despite increasing use of GPR in public health, current practice largely focuses on single-output prediction, calibration, or spatial risk mapping in limited policy spaces (Senanayake et al., 2016; Zimmer & Yaesoubi, 2020a). We lack spatially aware, multi-output metamodels that scale to exponentially many intervention combinations across counties and that guide simulation sampling over both locations and treatment condi-

tions while coherently propagating uncertainty to policy metrics. This work addresses that gap with a two-level metamodel in which the first level is a GPR and the second level is a response function that maps the GPR predictions to treatment effects, with a composite sequential design that allocates simulation runs across counties and treatment grids to maximize information gain.

1.1. Main related work

GPR has become a standard tool for emulating computationally intensive simulations in epidemiology and public health. By flexibly capturing nonlinear functions and providing calibrated uncertainty estimates, GPR offers a powerful surrogate for agent-based and compartmental epidemic models. Recent work demonstrates the potential of emulators to accelerate sensitivity analysis and calibration. 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).

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.

Another relevant strand of work centers on GPR, which has been widely adopted in healthcare and biomedical applications for flexible function approximation and uncertainty quantification. For example, GPR has been applied to ICU monitoring (Cheng et al., 2020) and pharmacology, as well as 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. In the epidemic modeling literature, outputs are still typically treated independently, missing opportunities to borrow statistical strength across correlated treatment-response coefficients. Finally, 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 both 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 treatment combinations across heterogeneous counties. Our work addresses this gap by developing a two-level metamodeling framework that

combines GPR with a response function and introduces a composite 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 that integrates spatial GPR, linear outcome models, and a two-stage sequential design strategy for selective simulation. Our approach enables the scalable approximation of simulation outcomes over a large geographic region, such as Pennsylvania, and readily generalizes to all U.S. counties with hundreds of treatment-condition combinations.

(1) GPR-Based Modeling of Spatially Varying Response-Function Coefficients We extend traditional GPR metamodeling to a spatially structured, multi-output setting that captures demographic heterogeneity and geographic continuity across counties in the contiguous U.S. Each county is encoded using its centroid coordinates and a set of socio-dynamic features, such as income, racial composition, and population density, forming a high-dimensional input space. The model outputs are the coefficients of a response function, which links treatment inputs to 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 at a single point in time. This formulation enables us to jointly learn the three response function parameters (μ_0, μ_1, μ_2) representing intercept, coefficient of first factor, and coefficient of second factor, under a shared spatial kernel structure, allowing for statistical borrowing of strength across locations and outputs.

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-demographic 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 and a periodic kernel, allowing the model to represent both smooth spatial variation and repeating attenuation patterns across regions.

(2) Composite sequential design for simulation-guided learning To overcome the computational burden of running more than one million simulations across C counties (e.g., 67 counties in Pennsylvania) and ℓ^J treatment conditions where ℓ denotes the number of levels per intervention and J denotes the number of interventions, we introduce a composite sequential design procedure. The design proceeds in two stages. In the first stage, we select which counties to simulate based on posterior uncertainty from 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 each selected county, we choose the most informative treatment condition based on the predictive variance of the outcome. This hierarchical design enables us to efficiently target simulations to regions where they are most needed, thereby accelerating

convergence while maintaining model accuracy.

(3) Spatial kernel design and feature selection. We construct a custom multi-component kernel for the GPR to capture county-level heterogeneity using spatial and contextual features. The kernel combines four Radial Basis Function (RBF) components and one periodic term. Each RBF is tied to a specific dimension, centroid coordinates, median income, population density, and Black population percentage, selected through a greedy forward selection strategy that evaluates cross-validated MSE. This kernel structure enables the GPR to learn smooth yet structured correlations across counties, improving predictive accuracy while maintaining interpretability.

This paper is organized into five sections. Section 2 introduces the two-level meta-model that combines a multi-output Gaussian process with a response function. Section 3 describes the composite sequential design framework, detailing how county- and treatment-condition sampling are integrated under a unified two-level procedure. Section 4 examines model, kernel, and sample complexity, highlighting the trade-offs between accuracy and efficiency. Finally, Section 5 concludes with a summary of findings and implications for policy evaluation.

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. Here, subgroups may represent geographic units (counties, regions), demographic strata (age or risk groups), or organizational entities (facilities, service areas). 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 unsimulated conditions, and (3) quantify predictive uncertainty to guide sequential allocation of simulation effort.

2.1. Two-level Metamodel: GPR and Response Surface Modeling

A conventional approach to metamodeling would employ a Gaussian Process 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 the proposed two-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), we model the opioid overdose death rate, measured as deaths per 100,000 people, using the following linear response function:

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

We refer to the response function $z(n, b \mid c)$ as the outcome level of the metamodel, which maps treatment condition (n, b) to predicted overdose mortality for each county c. In this formulation, n and b denote the levels of naloxone and buprenorphine, and $\mathbf{x}_c$ is the feature vector 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 are of three GPRs over spatial and socio-economic covariates of the counties denoted by $\mathbf{x}_c$.

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 two-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

The first-stage component of our two-level metamodel acts at the *contextual level* by modeling how the response-function coefficients depend on county-specific spatial and socio-economic features. In the first level, we employ a GPR to learn a vector-valued function that maps a county's feature vector $\mathbf{x}_c$ to the coefficients of the response function. Specifically, $\boldsymbol{\mu}(\mathbf{x}_c)$ represents the mean functions of the GPR derived from $\mathcal{GP}(\boldsymbol{\mu}(\mathbf{x}_c), \mathbf{K}(\mathbf{x}_c, \mathbf{x}_{c'}))$ where $\mathbf{K}(\mathbf{x}_c, \mathbf{x}_{c'})$ denotes the kernel function that encodes similarity between counties based on their spatial and socio-economic features, evaluated at $\mathbf{x}_c$, the vector of evaluated mean is expressed as:

$$\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c),\ \mu_n(\mathbf{x}_c),\ \mu_b(\mathbf{x}_c)]^\top.$$

To train the contextual GPR, 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,

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

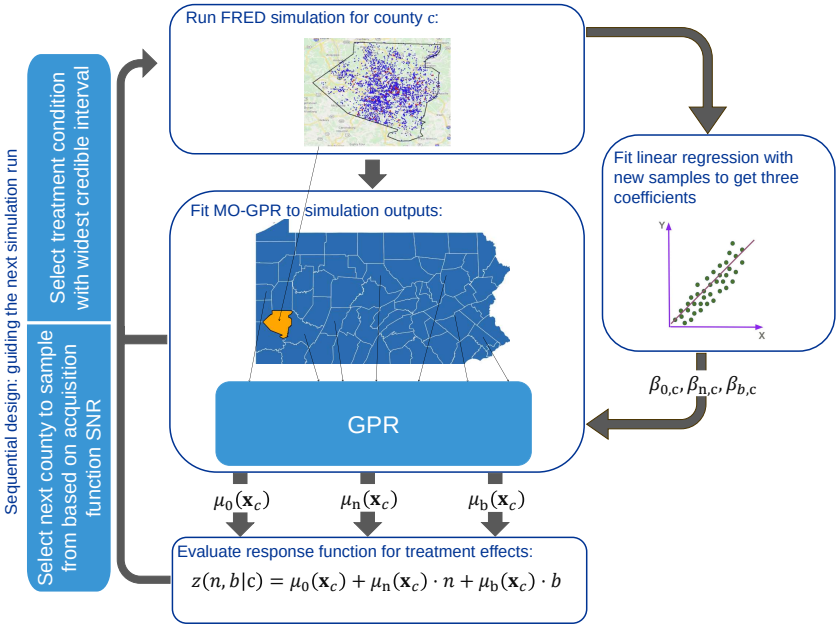


Figure 1: Overview of the two-level metamodeling framework. County-level outcomes are generated from FRED simulations (top), and a GPR is fit to summarize these outputs across counties (middle). The fitted GPR provides posterior distributions over the parameters of a linear response function, which is then used to evaluate treatment effects under any combination of naloxone (n) and buprenorphine (b) levels (bottom). Sequential design guides the selection of counties and treatment conditions for subsequent simulation runs.

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. Combining this likelihood with the GPR prior using Bayes' rule yields 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.

We adopt an independent-output formulation, in which each component output $\mu_m(\mathbf{x}_c)$ is modeled by its own Gaussian process over the same spatial–socioeconomic feature space. Collectively, these independent GPRs define a vector-valued prior over the response-function parameters,

$$\mathbf{f}(\mathbf{x}_c) = \left[f_0(\mathbf{x}_c),\ f_n(\mathbf{x}_c),\ f_b(\mathbf{x}_c)\right]^\top \sim \mathcal{GP}(\boldsymbol{\mu}(\mathbf{x}_c), \mathbf{K}(\mathbf{x}_c, \mathbf{x}_{c'}))\,, \tag{2}$$

where the kernel matrix $\mathbf{K}(\mathbf{x}_c, \mathbf{x}_{c'}) = diag[k_0, k_n, k_b]$ is block diagonal matrix and zeros elsewhere, reflecting the absence of cross-output covariance. Each diagonal kernel $k(\cdot, \cdot)$ may itself be constructed as a composition of kernels, for example through addition to capture independent effects or multiplication to encode interactions among input features.

The regression coefficients obtained from simulation runs are treated as noisy observations of these latent functions. For county c, the observation model is $\mathbf{y}_c \mid \mathbf{f}(\mathbf{x}_c) \sim \mathcal{N}(\mathbf{f}(\mathbf{x}_c), \boldsymbol{\Sigma}(\mathbf{x}_c))$, where $\boldsymbol{\Sigma}(\mathbf{x}_c)$ is a diagonal covariance matrix whose entries $\sigma_m^2(\mathbf{x}_c)$,

where $m \in \{0, n, b\}$, capture heteroscedastic uncertainty arising from finite simulation replicates and regression variability.

The heteroscedastic observation variances in the GPR are inherited directly from the regression uncertainty. 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, $\sigma_m^2(\mathbf{x}_c) \quad \propto \quad \mathrm{Var}(\widehat{\beta}_{m,c}) \,/\, R_c$, 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 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 composite 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. We use the signal-to-noise ratio (SNR) as our acquisition function. 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 $\boldsymbol{\mu}(\boldsymbol{x_c}) = [\mu_0(\mathbf{x}_c), \mu_n(\mathbf{x}_c), \mu_b(\mathbf{x}_c)]^\top$ and the associated covariance matrix $\boldsymbol{\Sigma}(\mathbf{x}_c)$. In our sequential design framework, the acquisition function is calculated directly from the posterior mean and covariance.

To compute a single acquisition value from this multi-output posterior, we em-

ploy 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. The SNR acquisition function is then defined as:

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

Sequential Sampling Procedure. At iteration t, the GPR posterior for county c provides $\boldsymbol{\mu}_t(\mathbf{x}_c)$ and $\boldsymbol{\Sigma}_t(\mathbf{x}_c)$, which 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 $c^* = \arg\max_{c \in \mathcal{C}} \alpha_{\text{SNR}}(c)$.

Based on this selection, we simulate S new county-condition pairs at c^* (see Section 3.2) and obtain the corresponding simulation outcomes. The outcomes from these runs are first used to fit a linear regression model that models the relationship between overdose deaths and intervention dispensing rates, 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.

This loop continues until a stopping criterion is met, such as when the variance becomes less than a specific value. In our implementation, the kernel encodes spatial and demographic similarity via a combination of multiple Radial Basis Functions These kernels guide both the similarity structure and uncertainty estimation, thereby influencing the sequential sampling trajectory.

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

```
Initialize GPR prior with spatial–socio-demographic kernel
Sample initial batch for each county to create dataset
for t = 1 to T do
    Update GPR posterior μ_t(x_c), Σ_t(x_c) using training dataset
    Compute scalarized mean and variance for each county c ∈ C:
        μ_t = w^⊤ μ_t(x_c)
        σ_t = √(w^⊤ Σ_t(x_c) w)
    Compute α_SNR(c)
    Select county c* = arg max α_SNR(c)
    Run simulation at (c*, (n*, b*)) (where (n*, b*), from Section 3.2)
    Update training dataset with new simulation outcomes
end for
```

3.2. Second-Stage Sequential Design for Sampling Treatment Conditions

At each iteration, the first-stage sequential design identifies a county for additional sampling, represented by its feature vector $\mathbf{x}_c$. However, evaluating all $\ell \times \ell$ treatment combinations for the selected county is computationally inefficient. The second stage of the sequential design helps us direct simulation effort toward the treatment con-

dition 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 given by $\mathbf{f}(\mathbf{x_c}) = \left[f_0(\mathbf{x}_c),\ f_n(\mathbf{x}_c),\ f_b(\mathbf{x}_c)\right]^\top$ in equation 2. 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) and 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) and 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:

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

with interval width

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

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

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

Algorithm 2 Second-Stage Sequential Design for Sampling Treatment Conditions

```
for each selected county feature vector x_c do
    Draw S posterior samples f^(s)(x_c) = [f_0^(s)(x_c), f_n^(s)(x_c), f_b^(s)(x_c)]^T
    for each treatment condition (n, b) do
        Compute predicted outcomes ζ^(s)(n, b | c) according to equation (3)
        Estimate 95% credible interval CI_95(n, b | c) using equation (4)
        Compute width(CI_95(n, b | c)) using equation (5)
    end for
    Select condition (n*, b*) = arg max_(n,b) width(CI_95(n, b | c))
    Run simulation at (n*, b*) and augment the training data
end for
```

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. Two-level Metamodel Workflow and Simulation Output Integration

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 this first level, the GPR is fit and updated using BoTorch's support for composite kernel structures and efficient batched posterior inference. As new county–condition observations are incorporated, the GPR posterior is recomputed, yielding updated mean vector $\boldsymbol{\mu}(\mathbf{x}_c)$.

where each element represents the posterior mean of the corresponding response-function parameter for county 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. The chosen condition is executed, the new outcome is appended to the training set, and the GPR is updated again using the same BoTorch implementation. This loop continues until the sample budget is exhausted or the error stabilizes. A complete summary of notation used in this workflow is provided in Appendix A.

Algorithm 3 GPR-RF Metamodeling Workflow

Initialize with a small set of county–treatment simulations
repeat
Stage 1, model update: Fit or update the BoTorch GPR on the current training data
Stage 1, county selection: Evaluate the SNR acquisition function over candidate counties, select the county c*
Stage 2, posterior sampling: For county with vector $\mathbf{x}_{c^*}$, draw posterior S samples of $[f_0(\mathbf{x}_{c^*}), f_n(\mathbf{x}_{c^*}), f_b(\mathbf{x}_{c^*})]$ from the GPR
Stage 2, treatment selection: For each treatment pair (n, b), compute the empirical 95% credible interval; select (n^*, b^*) with the widest interval
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 $\boldsymbol{\beta}(c^*)$
Simulation and augmentation: Run the simulation at (c^*, n^*, b^*) for q replications and append the new observation to the training set
until the sample budget is reached or the relative error stabilizes

4. Implementation and Numerical Results

We calibrated our OUD model to reproduce these county-level opioid mortality patterns. Six counties, Allegheny, Philadelphia, Dauphin, Erie, Columbia, and Clearfield, were calibrated using incremental mixture importance sampling. These counties were then used as prototypes to generalize model parameters across the remaining Pennsylvania counties via a feature-based matching procedure. The matching transfers calibrated model parameters governing disease 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 implementation details of the FRED simulation platform and the OUD model structure are provided in Appendix C, and the computational implementation is described in Appendix D.

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. For reference, training the full two-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, 2.25 GHz), and 12 GB of RAM, highlighting the practical computational efficiency of the proposed approach relative to exhaustive simulation.

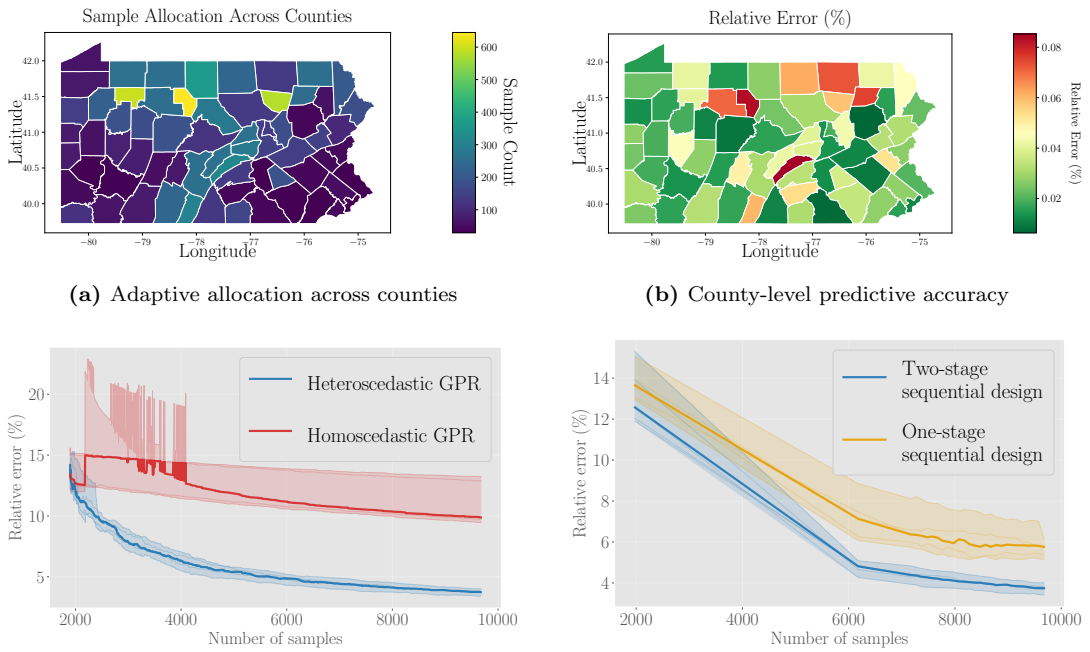
Model accuracy is evaluated on a held-out test set of simulation replications. For each county and treatment condition, test outcomes are computed by averaging across replications, and predictive performance is quantified using relative error and mean squared error with respect to these out-of-sample benchmarks.

4.1. Learning curves

We organize the subsequent empirical evaluation around three interrelated questions that govern the performance of the proposed two-level metamodel. First, how many simulation runs are required to achieve reliable predictive accuracy under 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 2a, 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 2b, with the corresponding sample allocation illustrated in Figure 2a. Compared with the approximately ~ 1.6 million simulations required for exhaustive enumeration, the two-level GPR–RF framework 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) Adaptive allocation across counties

(b) County-level predictive accuracy

(c) Noise modeling using heteroscedastic GPR

(d) Effect of sequential design on predictive accuracy

Figure 2: Empirical evaluation of sample efficiency, noise modeling, and sequential design strategies in the proposed two-level framework. Panels (a)–(b) illustrate how adaptive sampling concentrates simulation runs in difficult-to-learn counties and reduces county-level relative error. Panel (c) shows a learning curve demonstrating the benefits of heterosedastic noise modeling. Panel (d) 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.

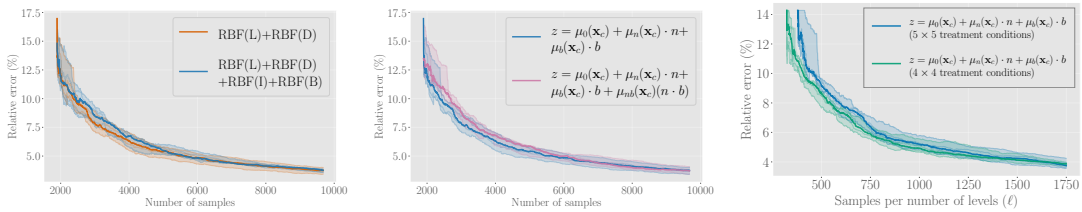
Explicitly modeling heterogeneous observation variance leads to substantially more stable and accurate learning behavior in the GPR. Under the heteroscedastic specification, relative error decreases smoothly as additional samples are incorporated, reflecting consistent posterior updating as uncertainty contracts in counties with increasing numbers of simulation replicates. In contrast, the homoscedastic formulation exhibits noticeably less stable learning, with oscillatory error trajectories at smaller sample sizes. 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 2c, 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.

After characterizing how different modeling choices influence learning behavior, another component of our analysis examines how the sequential design procedure impacts the rate at which the metamodel improves. Figure 2d 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 sam-

ples treatment conditions uniformly (orange curve). The county-condition 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.

Having established the benefits of modeling heteroscedastic uncertainty, we next examine the role of the response-function complexity in shaping metamodel performance. In particular, we contrast a main-effects response function with a more expressive specification that includes an interaction term, highlighting the trade-offs between sample efficiency, and representational capacity.

We compare the performance of a main-effects response function and an interaction-augmented specification within the metamodel framework. Figure 3b compares the learning curves of two response-function specifications evaluated under the same kernel design, $k(\mathbf{x}_c, \mathbf{x}_{c'}) = \mathrm{RBF}(L) + \mathrm{RBF}(D) + \mathrm{RBF}(I) + \mathrm{RBF}(B)$, where L, D, I, and B denote location, population density, income, and black population percentage. The main-effects specification, $z = \mu_0(\mathbf{x}_c) + \mu_n(\mathbf{x}_c)n + \mu_b(\mathbf{x}_c)b$, is contrasted with a more complex interaction specification, $z = \mu_0(\mathbf{x}_c)+\mu_n(\mathbf{x}_c)n+\mu_b(\mathbf{x}_c)b+\mu_{nb}(\mathbf{x}_c)(n\cdot b)$. 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 much larger simulation budget to be estimated reliably.



(a) Kernel **(b)** Response Function **(c)** Design Space

Figure 3: Analysis of sample efficiency and response-function complexity in the proposed framework. Panels (a) and (b) evaluate how kernel and model complexity configurations influence the rate of error decrease. Panel (c) compares outcome-level learning across intervention grids of different sizes (4×4 vs. 5×5), with samples normalized by the number of levels ℓ. This normalization highlights that, although per-level learning behavior is similar, the total number of simulations required to achieve a fixed prediction error increases with ℓ, consistent with the sample-complexity results in (Ahmed et al., 2024, Theorem 1).

In constructing our custom kernel, we incorporate four Radial basis function (RBF) kernels. In our kernel structure, each RBF component represents one specific contextual feature: the county's geographic location (via centroid coordinates), median household income, population density, and the percentage of Black residents. These socio-

demographic features are selected based on their relevance to opioid-related health disparities and intervention effectiveness. To assess the contribution of each feature–kernel combination, we compared the predictive MSE after training with N=10,000 samples.

We next examine how the choice of kernel influences metamodel performance. Figure 3a reports relative-error learning curves for two kernel specifications applied to the same response function. The simpler specification, $\mathrm{RBF}(L) + \mathrm{RBF}(D)$, captures spatial and demographic variation using only location (L) and population density (D). The more expressive specification, $\mathrm{RBF}(L)+\mathrm{RBF}(D)+\mathrm{RBF}(I)+\mathrm{RBF}(B)$, 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.

To assess outcome-level learning behavior under changes in the intervention design space, we evaluate the performance of the metamodel across treatment grids of different sizes. Figure 3c 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 $\sim$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.

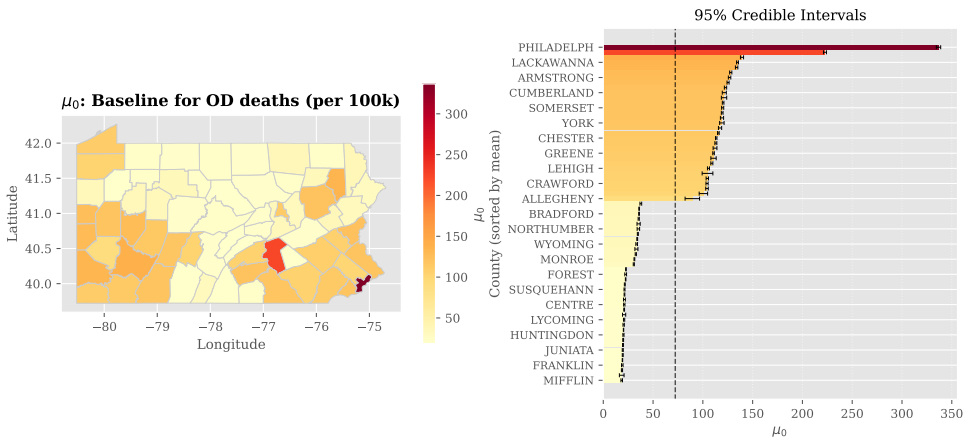
4.2. Localized Treatment Effects

Substantial heterogeneity is observed in both baseline overdose mortality and intervention response across Pennsylvania counties. Baseline levels differ markedly 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 4.

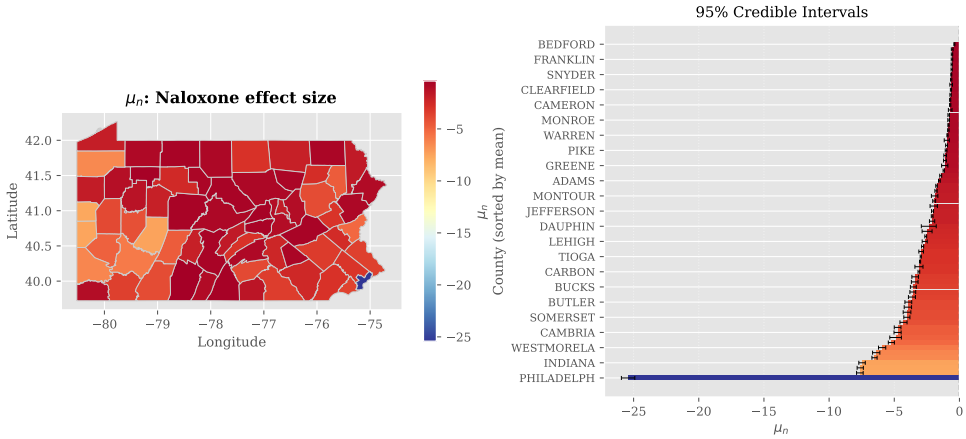
To better understand the spatially localized and heterogeneous treatment effects identified above, Table 1 reports posterior means and 95% credible intervals for the response-function coefficients in the six calibrated counties used as prototypes in the modeling framework. Philadelphia exhibits the highest 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.

From localized treatment effects to localized 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.

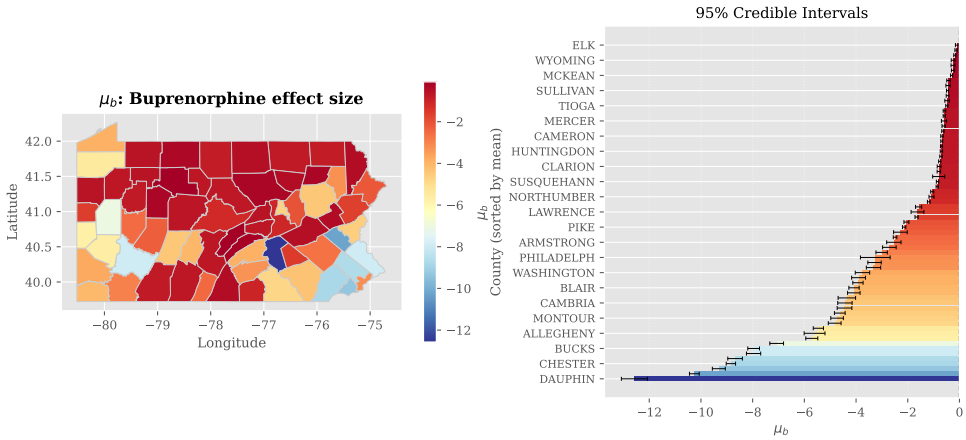
Recent simulation-based studies of opioid interventions find that the combined effects of harm-reduction strategies (i.e., naloxone distribution), medications for opioid use disorder (i.e., buprenorphine treatment), and diversion-based policies depend on local conditions, and that these interventions 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 risk and treatment responsiveness.



(a) Estimated Baseline overdose mortality (μ_0) per 100,000 people, with corresponding 95% credible intervals.



(b) County-specific naloxone effect sizes (μ_n) and associated 95% credible intervals.



(c) County-specific buprenorphine effect sizes (μ_b) and associated 95% credible intervals.

Figure 4: Posterior summaries of the GPR-estimated response-function coefficients across PA 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.

Table 1.: 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]

5. Conclusion

This study presents a novel two-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-demographic features, learns 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 two-level metamodel, when coupled with our 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. The policy space considered includes six interventions, each varied across seven levels, yielding thousands of possible treatment configurations (7^6) that would be computationally inefficient to evaluate directly. 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 prioritizes numerical stability and avoids overparameterization in high-dimensional feature spaces. 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 predictive performance in regions where treatment effects exhibit shared spatial or socio-demographic 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-demographic features, but it ignores temporal evolution in treatment effects and epidemic dynamics. In reality, both epidemic outcomes and intervention impacts evolve over time, shaped by policy shifts and emerging drug trends. 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-demographic features but also on time.

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, naloxone, 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`.

Funding

This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC and VIZ clusters, which are supported by NIH award number S10OD028483.

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).

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, *4*(3), 195–266.

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.

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.

Binois, M., Gramacy, R. B., & Ludkovski, M. (2018). Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, *27*(4), 808–821.

Bosetti, P., Poletti, P., Stella, M., Lepri, B., Merler, S., & De Domenico, M. (2020). Heterogeneity in social and epidemiological factors determines the risk of measles outbreaks. *Proceedings of the National Academy of Sciences*, *117*(48), 30118–30125.

Box, G., & Wilson, K. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(1), 1–45.

Box, G. E., & Draper, N. R. (1975). Robust designs. *Biometrika*, *62*(2), 347–352.

Box, G. E., & Hunter, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *The Annals of Mathematical Statistics*, 195–241.

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])

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, 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.

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.

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.

Georganos, S., & Kalogirou, S. (2022). A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS International Journal of Geo-Information*, *11*(9), 471.

Giabbanelli, P., & Crutzen, R. (2013). An agent-based social network model of binge drinking among dutch adults. *Journal of Artificial Societies and Social Simulation*, *16*(2), 10.

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

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.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini Jr, I. M., Cummings, D. A., Lewis, B., ... others (2008). Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, *105*(12), 4639–4644.

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.

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).

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.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

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

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments.* John Wiley & Sons.

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.

Salle, I., & Yıldızoğlu, M. (2014). Efficient sampling and meta-modeling for computational economic models. *Computational Economics*, *44*(4), 507–536.

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.

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).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

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*.

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. Notation

This appendix summarizes the notation used throughout 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.

Table A1.: Summary of notation used in the paper.

Symbol	Description
J	Number of interventions (factors)
ℓ	Number of dispensing rate levels per intervention
t	Iteration index in the sequential design algorithm
q	Number of replications per selected simulation condition
$(n, b) \in \mathcal{T}$	Treatment condition defined by naloxone level n and buprenorphine level b
$\mathcal{D}$	Training set
$\mathcal{C}$, $\lvert\mathcal{C}\rvert$	Set of counties, number of counties
$\mathbf{x}_c \in \mathbb{R}^d$	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
$\boldsymbol{\mu}(\mathbf{x}_c) = [\mu_0(\mathbf{x}_c), \mu_n(\mathbf{x}_c), \mu_b(\mathbf{x}_c)]^\top$	GPR posterior mean of coefficients
$\boldsymbol{\Sigma}(\mathbf{x}_c)$	GPR posterior covariance of coefficients
$z(n, b \mid c)$	Response function prediction under (n, b) for county c
$\zeta^{(s)}(n, b \mid c)$	Posterior sample of $z(n, b \mid c)$ (index s)

Appendix B. 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. Most recently, Ahmed et al. (Ahmed et al., 2023b) demonstrate that GPR regression can capture distinct epidemic trajectories even when initial parameters such as location, population, and contact rates are identical. Ahmed et al. (Ahmed et al., 2023a) demonstrate that a regression-based greedy sampling strategy, which allocates more simulation runs to treatment conditions with wider credible intervals, can estimate treatment effects with accuracy comparable to equal-allocation strategies while requiring substantially

fewer samples. This finding highlights the importance of informed sequential sampling in large-scale simulation studies.

Kriging is a closely related method originally developed for spatial interpolation in geostatistics. The term "Kriging" was coined in honor of Danie Krige, who first introduced the technique for estimating values at unsampled locations based on observations from sampled points (Forrester et al., 2008; Matheron, 1963). In modern applications, particularly within engineering and machine learning, Kriging is often considered equivalent to GPR, which emerged later and has been widely adopted in surrogate modeling (Gramacy, 2020). Both methods share the underlying probabilistic framework of defining a distribution over functions and making predictions with associated uncertainty. Kriging relies on a variogram, a function that models spatial dependence, to quantify the relationship between locations. This variogram serves a similar role to the kernel function in GPR, and is typically fitted to the observed data to guide the interpolation process (Moraga, 2023). Georganos et al. (Georganos & Kalogirou, 2022) introduce a spatially weighted formulation of Geographical Random Forests that incorporates local spatial structure directly into the model and greatly improves predictive accuracy for heterogeneous geographic data. Their approach also provides substantial computational savings through an efficient implementation, enabling scalable spatial machine-learning analysis over large regions.

The basic setting of the multi-armed bandit (MAB) problem involves K arms, each associated with an unknown probability distribution. At each iteration, an agent selects one arm and receives a reward, which is drawn from the selected arm's distribution and assumed to be independent of previous selections. The objective of the agent is to maximize cumulative reward over time by learning to favor arms with higher expected returns. This problem naturally introduces the exploration, exploitation tradeoff, where the agent must decide whether to continue sampling from arms with known rewards or to explore less-sampled arms that may yield better outcomes (Sutton & Barto, 2018).

Response Surface Methodology (RSM) is a collection of statistical and mathematical techniques used to develop, improve, and optimize processes. It investigates the relationship between explanatory variables and one or more response variables, and has applications in both the design of new products and the refinement of existing ones. RSM is particularly common in industrial settings, especially when multiple input variables influence the performance or quality characteristics of a product or process. In such applications, the outputs, referred to as response variables, are modeled as functions of controllable input variables. The term "response surface methodology" originated from scientific experimentation, for example, in chemical engineering, where the goal is to predict the outcome of a chemical process based on specific combinations of input elements. A core feature of RSM problems is that the true response surface is unknown, and the primary objective is to identify input settings that optimize the desired response (Myers et al., 2016). RSM was first introduced by (G. Box & Wilson, 1951), and subsequently extended by (G. E. Box & Hunter, 1957; G. E. Box & Draper, 1975), who examined the robustness of experimental designs and proposed multi-level approaches to account for model uncertainty.

Appendix C. FRED agent-based modeling and calibration

The Framework for Reconstructing Epidemiological Dynamics (FRED) is an open-source, agent-based modeling platform developed by the Public Health Dynamics Lab-

oratory (PHDL) at the University of Pittsburgh to simulate the spread of infectious diseases. 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.

C.1. 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) analyzed 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 enters 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 CDC data through model calibration to accurately reflect real-world dynamics. 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 through exposure. From either of these states, the agent can then transition into the OUD state, which may lead to overdose death (ODD), other cause 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 of the OUD simulation 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 nonuser 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. In this study, we use fixed intervention levels for simplicity, while future work will consider a more detailed intervention model with time-varying or population-specific strategies.

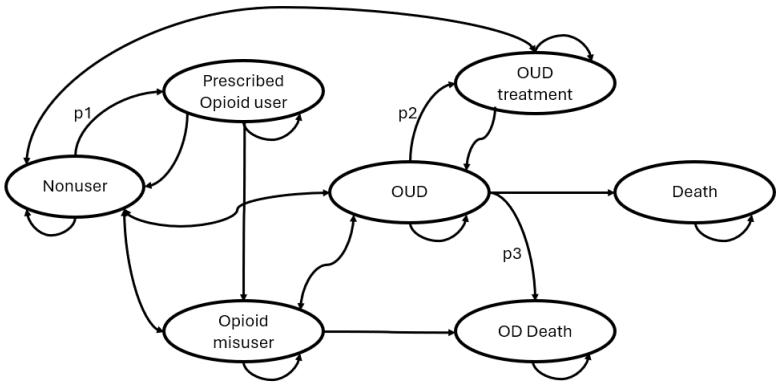


Figure C1: State transition diagram for the OUD model. Transition probabilities p_1, p_2, and p_3 are defined by Equations (C1)-(C3), respectively.

C.2. Model Calibration for Selected Counties with a Bayesian Importance Sampling Method

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 counts. 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) 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, medium, and small. Each pair is designed 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. We emphasize that calibration is an expert-driven and resource-intensive process. It involves manual validation, checking simulation trajectories, tuning parameter bounds, and refining the objective function. 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.

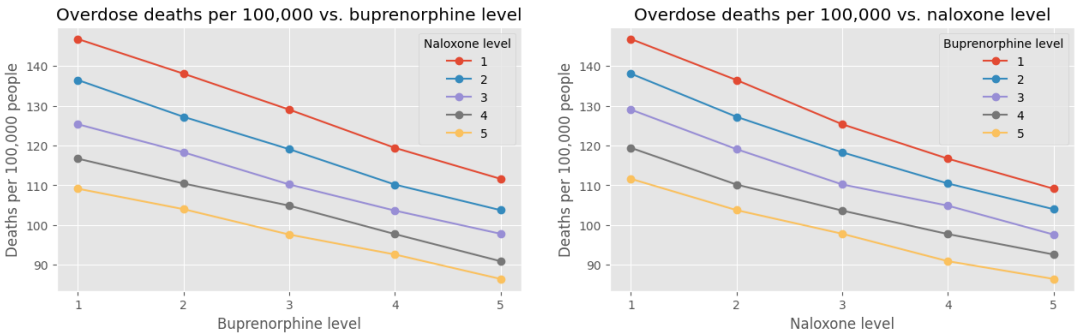
C.3. Model Selection for the Response Function

To assess whether an interaction is needed between naloxone and buprenorphine, we use two factorial plots (Fig. C2): 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. We therefore adopt a simple main-effects model.

Appendix D. Implementation Details and Computational Constraints

As we mentioned in the previous section, calibrating agent-based models such as FRED for each individual county is computationally expensive and methodologically intensive. 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.

For each county, we construct a slope–based summary vector that captures both the level and temporal trend of opioid indicators over 2015–2019. Denote by $\mathbf{v}_{c\tau} =$



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

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

Figure C2: Factorial plots over the 5×5 intervention grid. The absence of line intersections and the near-parallel trends in both figures provide no visual evidence of interaction between naloxone and buprenorphine. A simple main-effects model, $z(n, b \mid c) = \mu_0(\mathbf{x}_c) + \mu_n(\mathbf{x}_c) \cdot n + \mu_b(\mathbf{x}_c) \cdot b$, is therefore sufficient for interpretation and outcome-level metamodeling.

$(v_{c\tau}^{\text{opioid}}, v_{c\tau}^{\text{nalox}}, v_{c\tau}^{\text{bup}}, v_{c\tau}^{\text{fent}})$ the standardized dispensing–rate vector for county c in year τ, and let $d_{c\tau}$ be the standardized overdose–death rate. A simple linear regression of each time series on calendar year provides:

$$\text{AvgDeath}_c = \frac{1}{T}\sum_{\tau=1}^{T} d_{c\tau}, \qquad \text{DeathTrend}_c = \widehat{\text{slope}}\{d_{c\tau}\},$$

$$\text{FeatureMag}_c = \left\|\widehat{\text{slope}}\{\mathbf{v}_{c\tau}\}\right\|_2, \qquad \text{Population}_c = \text{pop}_{c,latest},$$

where $\widehat{\text{slope}}\{\cdot\}$ denotes the ordinary–least–squares slope across the five annual observations ($\tau = 5$). Each of the four features is then z-score standardized across counties prior to distance computation, producing a standardized feature vector $\mathbf{u}_c \in \mathbb{R}^4$ whose components are on comparable scales for that county, ensuring that subsequent distance calculations reflect relative, not absolute, magnitudes. Given the six calibrated prototype counties $\mathcal{P}$, every non-calibrated county c is matched to its nearest prototype $p \in \mathcal{P}$ in this feature space:

$$d_{cp} = \|\mathbf{u}_c - \mathbf{u}_p\|_2, \quad p \in \mathcal{P}, \qquad \text{Prototype}(c) = \arg\min_{p \in \mathcal{P}} d_{cp}.$$

Because the standardization is performed across counties, the Euclidean distance d_{cp} highlights similarities in relative feature profiles rather than absolute magnitudes. The resulting mapping enables the transfer of calibrated OUD parameters from the six prototypes to all remaining counties, thereby eliminating the need for hand-tuning calibration for the 61 additional counties and substantially reducing both computational

and manual effort.

D.1. Scalable Simulation Execution and Cluster 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.