Addressing positivity violations in causal effect estimation using Gaussian process priors

Angela Yaqian Zhu | Nandita Mitra | Jason Roy

1Janssen Research & Development, Johnson & Johnson, Raritan, New Jersey, 2Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 3Department of Biostatistics and Epidemiology, School of Public Health, Rutgers University, Piscataway, New Jersey,

Correspondence
Angela Yaqian Zhu, Janssen Research & Development, Johnson & Johnson, Raritan, NJ 08869, USA.
Email: azhu147@its.jnj.com

In observational studies, causal inference relies on several key identifying assumptions. One identifiability condition is the positivity assumption, which requires the probability of treatment be bounded away from 0 and 1. That is, for every covariate combination, it should be possible to observe both treated and control subjects the covariate distributions should overlap between treatment arms. If the positivity assumption is violated, population-level causal inference necessarily involves some extrapolation. Ideally, a greater amount of uncertainty about the causal effect estimate should be reflected in such situations. With that goal in mind, we construct a Gaussian process model for estimating treatment effects in the presence of practical violations of positivity. Advantages of our method include minimal distributional assumptions, a cohesive model for estimating treatment effects, and more uncertainty associated with areas in the covariate space where there is less overlap. We assess the performance of our approach with respect to bias and efficiency using simulation studies. The method is then applied to a study of critically ill female patients to examine the effect of undergoing right heart catheterization.

KEYWORDS
Bayesian nonparametrics, causal inference, extrapolation, Metropolis within Gibbs, overlap, population average treatment effect

1 INTRODUCTION

Researchers often aim to infer the causal effects of a treatment on a population of interest from observational studies. Identification of causal effects from observational data relies on assumptions including no unmeasured confounding, which holds when treatment assignment is random (that is, independent of potential outcomes) given measured confounders.1 If the treatment received depends on observed covariates, then the distribution of these covariates is expected to differ by treatment group. This raises concerns about violations of a second identifying assumption called positivity. Positivity assumes that there is a nonzero probability of receiving treatment for all individuals. If there is a subpopulation defined by covariates for which one of the treatments is not observed, causal contrasts for that subgroup cannot be identified without further assumptions.2,3

Abbreviations: ATE, average treatment effect; BART, Bayesian additive regression trees; GP, Gaussian process.

Nandita Mitra and Jason Roy are co-senior authors.
Theoretical (or structural) violation of the positivity assumption occurs when a subpopulation of individuals have zero probability of receiving at least one of the treatments, so that even if we let the sample size go to infinity, we would still never observe all treatment values. This can happen, for example, when treatment is contraindicated in a certain subgroup of patients because of their age, comorbidities, and family history of disease. D’Amour et al elaborate on this type of violation in the context of high-dimensional covariates. On the other hand, practical (also called random) violation of positivity arises when, in a given observational dataset, a subpopulation is not observed to receive a particular treatment by chance. For example, suppose in a sample, no males between the ages of 35 to 45 receive treatment A purely by chance. In reality, their probability of receiving treatment A may be small but not zero. In this case, we will not be able to learn about the treatment effect of A in this subpopulation of men without making additional modeling assumptions. We expect practical positivity violations to arise in clinical data, especially when there are a large number of covariates. However, we could potentially learn about these nonoverlap regions using modeling. For instance, if we are willing to assume an additive linear regression model, we could learn about males treated with treatment A via linear interpolation or extrapolation from younger or older men who received treatment A. The disadvantage is that we would need to rely on strong parametric assumptions. Further, these approaches may underestimate the degree of uncertainty that would be expected in data-sparse regions.

Trimming approaches are commonly used to address positivity violations and are discussed in Petersen et al. Crump et al. propose a method that removes (trims) subjects whose propensity scores are outside specified bounds and calculates a minimum variance estimate on the remaining subsample. This approach requires correct specification of the propensity score model and may result in a final sample that is a smaller subsample of the original study population. Others (Rosenbaum and Visconti & Zubizarreta) have suggested matching treated and control subjects on covariates or propensity scores; however, external validity may be diminished due to matching because the target population of interest will have changed to that of the matched population. Hill and Su define nonoverlap as the set of subjects whose estimated individual causal effects have corresponding variances that are greater than specified upper bounds. However, these upper bound cut-offs may rely on user specifications and may not adequately reflect the amount of data sparsity. Ghosh characterizes multivariate covariate overlap using convex hulls to determine positivity violations. This overlap subset is termed “the margin” and is determined using a propensity score model; subjects who are outside the margin are trimmed. A disadvantage to these trimming procedures is that by discarding subjects, the target of inference may shift to a resulting subpopulation that may not represent the original population of interest.

An alternative to removing subjects entirely is to downweight subjects in regions with less overlap. In that spirit, Li et al. proposed estimating causal effects using overlap weights. However, overlap weights involve propensity score estimation and place more emphasis on those with higher probabilities of receiving either treatment. Although trimming or weighting approaches may be appropriate for structural violations, they fall short for practical violations of the positivity assumption. In this practical violations setting, we expect covariates of subjects who violate this assumption to not be too far from the area of the covariate space where there is complete overlap. In other words, we consider the setting where our study objectives aim to make inference for a population but, by chance, the available sample of individuals does not include certain individual characteristics for all treatment levels. For example, in studies that seek to inform public policy, population-level inference is desired since changes in policy will affect the general population. However, in a particular dataset there could be some nonoverlap by chance. Or, consider a study comparing the safety and effectiveness of drug A vs drug B. Suppose there happens to be no Hispanic individuals over age 60 who take drug A, but we expect that if we had a larger study that cell would no longer be empty. We do not want to exclude Hispanic individuals over age 60 from our analysis, but we do want our inferential procedures to account for the fact that contributions to the overall treatment effect estimate from this subpopulation will involve extra uncertainty due to extrapolation. In situations when there is an intended patient population for a treatment or intervention, but this population may not be entirely reflected in the available treatment group data, methods for population-level estimation are needed. The objective in these studies is often to obtain inference that preserves the original population, ensuring results may be generalized accordingly. When there are practical but not structural violations, interest generally centers on understanding the treatment effects for the entire population. Approaches that account for positivity violations while also targeting a causal estimand for the whole population are, therefore, of most interest.

There has been some recent work on methods that are based on extrapolating information from overlap regions to nonoverlap regions in a way that preserves the intended target of inference. For instance, Nethery et al. propose a method for estimating a causal effect on the entire population using extrapolation. Their definition of the overlap region relies on two user-specified parameters that indicate the desired extent of closeness between treatment groups based on subjects’ estimated propensity scores. Choices regarding propensity score model specification and user inputs influence whether
a subject is included in the overlap region. Having a fixed region means that uncertainty around subjects’ inclusion in the overlap or nonoverlap region is ignored. In this approach, Bayesian additive regression tree (BART) models are used to estimate causal effects for the overlap regions and then in a subsequent stage, spline (SPL) models extrapolate those trends to subjects in the nonoverlap region. These choices in modeling mean that prior information on the treatment effect cannot be directly utilized. Lastly, although they account for the greater uncertainty in areas of nonoverlap, their proposed variance inflation strategy results in over-coverage as seen in their simulation studies.

To address some of the above limitations of existing approaches, we propose a Gaussian process modeling approach for estimating average treatment effects in a way that preserves the original target of inference when there are practical positivity violations. Our method contributes several advances to the current literature. First, because we use a nonparametric prior distribution, we avoid making parametric modeling assumptions. Further, the prior does not rely on user-specified parameters nor cut-offs to define regions of overlap. The amount of nonoverlap is accounted for in the covariance functions of the Gaussian process as the distances of a subject’s covariate values from those of individuals in the other treatment group. This provides us with a sense for nonoverlap that is data driven. We also avoid the need to construct explicit overlap and nonoverlap groups, allowing covariate distance and positivity violations to be considered in a more continuous fashion. Importantly, the further subjects are from each other in terms of their covariate values, the larger the variances of their individual causal effects, which reflects the underlying point that there should be greater uncertainty around estimated causal effects when there is less overlap.

The remainder of the article is organized as follows. In Section 2, we formulate the Gaussian process model and present the Bayesian inferential framework with its priors, likelihood, and posteriors. In Section 3, we conduct simulation studies to assess the performance of our method compared to other current approaches. We then apply our approach to data from an observational study of right heart catheterization in female patients in Section 4. Section 5 provides a discussion of results and offers concluding remarks.

2 | THE GAUSSIAN PROCESS MODEL AND POSTERIOR COMPUTATION

2.1 | Notation and framework for causal effect estimation

Here, we use the potential outcomes framework for estimating causal effects. Suppose there are \( n \) i.i.d. observations from a population. For each subject \( i \) in the sample, let \( A_i \) be the treatment assignment indicator with \( A_i = 1 \) if subject \( i \) receives the treatment of interest and \( A_i = 0 \) if subject \( i \) receives the control. For a dichotomous treatment, each subject \( i \) has two potential outcomes: \( Y_i(1) \), the outcome under treatment, and \( Y_i(0) \), the outcome under control. However, each subject may only receive one treatment in a study, and the observed outcome for subject \( i \), \( Y_i \), reflects the treatment received. Furthermore, define \( X_i \) to be a vector of \( p \) pre-treatment variables or covariates.

Our target parameter of interest is the mean difference in potential outcomes under treatment and under control, respectively, given by \( \Psi = E[Y(1) - Y(0)] \). This represents the average effect had everyone in the representative population been given treatment vs had everyone in the representative population been given control. Here we assume that there is superpopulation of units from which the study sample is drawn, consisting of individuals who are eligible for treatment. Due to the finite size of the study sample, positivity violations can occur when certain patient characteristics are not observed in the treated sample. Identifiability of this parameter rests on the following assumptions:

1. Consistency: \( Y = Y(a) \) whenever \( A = a \).
2. Ignorable treatment assignment: Conditional on covariates, treatment assignment is independent of the set of potential outcomes, \( A \perp \{Y(0), Y(1)\} | X \). This essentially says that all confounding may be accounted for.
3. Positivity: The probability of receiving either treatment given the covariates is nonzero, \( 0 < P(A = 1 | X) < 1 \).

2.2 | Gaussian process model

We assume the model for the observed continuous outcome \( Y \) given confounders \( X \) and treatment \( A \) has the following form:

\[
Y_i = \mu(X_i) + \Delta(X_i)A_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2).
\]
The function $\mu(X_i)$ represents the relationship between $X$ and $Y$ that is not part of the treatment effect; that is, it is the prognostic impact of covariates. The function $\Delta(X_i)$, which is a functional coefficient of $A_i$, can be thought of as representing conditional treatment effects, reflecting interactions between covariates and treatment. Under the causal assumptions described above, the average causal effect is just $\Psi = E(\Delta(X))$.

We treat the functional form of $\mu()$ and $\Delta()$ as unknown, and therefore need to specify priors for those functions. We assume independent Gaussian process (GP) priors for these functions. Specifically,

$$\mu(X) \sim GP(X\beta, \mathcal{K}_\mu),$$
$$\Delta(X) \sim GP(0, \mathcal{K}_\Delta).$$

The mean function in the prior for $\mu(X)$ centers this parameter on a linear model, $X\beta$, while the mean function for $\Delta(X)$ is zero to reflect the a priori belief of small heterogeneous treatment effects; that is, the treatment effect (both main effect and effect modification) is centered at 0. An advantage of GP priors is that we can center the priors on a parametric model. Essentially, the prior mean based on these functions is a linear model with no effect modification. Thus, when there is limited data, the outcome model will shrink towards this prior specification. An alternative is to center the GP prior for $\Delta(X)$ on a nonzero scalar, reflecting a priori a homogeneous causal effect. For instance, if $\Delta(X)$ had a GP prior with mean $\lambda$, then the treatment effects would shrink toward a common effect $\lambda$.

With the goal of having a noisier mean function when there is less overlap, we choose the squared exponential (SQEXP) form for the covariance functions. For subjects $i$ and $j$ with covariate vectors $X(i) = \{X_1(i), \ldots, X_p(i)\}$ and $X(j) = \{X_1(j), \ldots, X_p(j)\}$, respectively, the covariance function $\mathcal{K}_\Delta$ defines the $(i,j)$th element of the covariance matrix $K_\Delta$ to be

$$K_{\Delta,ij} = \mathcal{K}_\Delta(X(i),X(j)) = \eta^2_\Delta \exp \left\{ -\frac{1}{2} \sum_{p=1}^{P} \frac{\left[ X_p(i) - X_p(j) \right]^2}{l_\Delta^2} \right\}.$$

The covariance $\mathcal{K}_\mu$ has the same form, but with different parameters, $l_\mu$ and $\eta_\mu$. The reason that this particular covariance function is useful when there are practical violations of the positivity assumption is that the variability of the function increases as distance between covariates increases, which we later show using the posterior distribution of $\Delta$.

The hyperparameters $l$ and $\eta$ in the GP prior determine the shape and smoothness of functions defined by the prior distribution. The parameters $l_\mu$ and $l_\Delta$ are the length scales which characterize the extent to which $\mu$ and $\Delta$ function values change as the input changes.\(^{17}\) Small values correspond to more frequent changes in the parameter values for the same change in inputs $X$; that is, the distance in $X$ needed for the parameters to vary by an amount comparable to its range is smaller for small length scales. Larger values of this hyperparameter correspond to more smooth curves a priori. The parameters $\eta_\mu$ and $\eta_\Delta$ are the signal variances (output-scale amplitude), which control the range of the function values. For $\eta$ near 0, posterior mean estimates of the parameters $\mu$ and $\Delta$ tend to be close to each other with fewer fluctuations in the curve (closer to a straight line). At larger $\eta$ values, regions with nonoverlap will have more variability associated with the corresponding causal effect for a particular combination of covariate values.

### 2.2.1 Choice of kernel

The kernel or covariance function determines the types of statistical structures that may be captured by the GP model.\(^ {22}\) Our model is presented using the SQEXP kernel, resulting in functions that are infinitely differentiable to allow for smoothing. However, other stationary kernels, which only depend on the distance between two points, may also be appropriate since these methods all aim to capture the similarity in baseline characteristics of subjects.\(^ {23}\) Common covariance function specifications include the following.\(^ {16}\)

- **Rational quadratic:** $k(x, x') = \eta^2 \left( 1 + \frac{\|x-x'\|^2}{2\alpha^2} \right)^{-\alpha}$.

  This kernel is infinitely mean square differentiable and is a scale mixture of SQEXP kernels with different length-scales. The limit of this covariance function as $\alpha \to \infty$ is the SQEXP kernel.
• Matérn: \( k(x, x) = \eta^2 \frac{2^{1-v}}{\Gamma(v)} \left( \frac{\sqrt{2v|\mathbf{x}-\mathbf{x}'|}}{l} \right)^{v} K_v \left( \frac{\sqrt{2v|\mathbf{x}-\mathbf{x}'|}}{l} \right), \) where \( l \) and \( v \) are positive hyperparameters and \( K_v \) is the modified Bessel function.\(^{24,25}\)

This kernel also converges to the SQEXP kernel as \( v \to \infty.\)

• Exponential covariance function (or Ornstein-Uhlenbeck in the one-dimensional case): \( k(x, x') = \eta^2 \exp \left( -\frac{|x-x'|}{l} \right).\)\(^{26}\)

We consider these three specifications of the covariance and compare their performances to our proposed SQEXP kernel in our simulations. In all cases, the hyperparameters are given hyperpriors such that their values are sampled in each iteration of the Markov chain Monte Carlo (MCMC) chain to allow the data to influence their values. In practice, these kernels may be combined by addition or multiplication so that the resulting kernel may capture more complexities in the data.\(^{22}\)

### 2.3 Priors, likelihood, and posteriors

The outcome given treatment and confounders is distributed as \( Y \sim \text{MVN}(\mu + \Delta A, \sigma^2 I), \) which implies that the likelihood is

\[
p(y|\mu, \Delta, \sigma^2) \propto \sigma^{-2} \exp \left[ -\frac{1}{2} (y - (\mu + \Delta A))^T (\sigma^2 I)^{-1} (y - (\mu + \Delta A)) \right].
\]

Note that \( \mu = \mu(X) \) and \( \Delta = \Delta(X) \) for simplification of notation. Also, \( \Delta A = \begin{bmatrix} \Delta A_1 \\ \vdots \\ \Delta A_n \end{bmatrix} \).

We specify priors for the hyperparameters as follows. The vector of coefficients \( \beta \) is given a multivariate normal prior, \( p(\beta) \propto \text{MVN}(0, \sigma^2 I) \). This conjugate prior leads to a conditional posterior distribution for \( \beta \) that is also multivariate normal. The hyperparameters \( l \) and \( \eta \) are given gamma priors since their values need to be positive: \( p(l_{\mu}) \propto \text{gamma}(l_{\mu}|a_{l_{\mu}}, \beta_{l_{\mu}}), p(\eta_{\mu}) \propto \text{gamma}(\eta_{\mu}|a_{\eta_{\mu}}, \beta_{\eta_{\mu}}), p(l_{\Delta}) \propto \text{gamma}(l_{\Delta}|a_{l_{\Delta}}, \beta_{l_{\Delta}}), \) and \( p(\eta_{\Delta}) \propto \text{gamma}(\eta_{\Delta}|a_{\eta_{\Delta}}, \beta_{\eta_{\Delta}}) \). The hyperparameter \( \sigma^2 \) has an inverse-gamma prior, \( p(\sigma^2) \propto \text{Inv} - \text{gamma}(\sigma^2|a_{\sigma^2}, \beta_{\sigma^2}) \), which is a common prior for variance parameters. The joint prior is

\[
p(\mu, \beta, l_{\mu}, \eta_{\mu}, \Delta, l_{\Delta}, \eta_{\Delta}, \sigma^2) \propto p(\mu|l_{\mu}, \eta_{\mu}) p(\beta|l_{\mu}, \eta_{\mu}) p(l_{\mu}) p(\eta_{\mu}) p(\Delta|l_{\Delta}, \eta_{\Delta}) p(l_{\Delta}) p(\eta_{\Delta}) p(\sigma^2),
\]

which assumes a priori independence of the hyperparameters.

Then the joint posterior is

\[
p(\mu, \beta, l_{\mu}, \eta_{\mu}, \Delta, l_{\Delta}, \eta_{\Delta}, \sigma^2|Y) \propto p(y|\mu, \Delta, \sigma^2) p(\mu|l_{\mu}, \eta_{\mu}) p(\beta|l_{\mu}, \eta_{\mu}) p(l_{\mu}) p(\eta_{\mu}) p(\Delta|l_{\Delta}, \eta_{\Delta}) p(l_{\Delta}) p(\eta_{\Delta}) p(\sigma^2).
\]

### 2.3.1 Conditional posteriors

To ensure the Gaussian process priors for \( \mu \) and \( \Delta \) are not in too much disagreement with the data, we estimate hyperparameter values based on data and the posterior. That is, rather than choosing fixed values for the hyperparameters, \( l \) and \( \eta \), in the GP priors, we assign them gamma priors as specified in the previous subsection and use Metropolis-Hastings to update their values. These are integrated in a Metropolis within Gibbs algorithm to obtain posterior inference for \( \mu \) and \( \Delta \).\(^{27}\) The conditional distributions for \( \beta, \mu, \) and \( \Delta \) have analytical forms, so estimates of these parameters may be drawn directly in their corresponding Gibbs steps.\(^{26,29}\) Because these conditional distributions are needed in the algorithm and are specific to the form of our Gaussian process model, we present them here. Detailed derivations are provided in Appendix A in Data S1.

- \( \beta|\mu, l_{\mu}, \eta_{\mu}, X, y \sim \text{MVN}\left(\begin{bmatrix} X^T K^{-1}_{\mu} X + (\sigma^2_{\beta} I)^{-1} \end{bmatrix}^{-1} X^T K^{-1}_{\mu} \mu, \begin{bmatrix} X^T K^{-1}_{\mu} X + (\sigma^2_{\beta} I)^{-1} \end{bmatrix}^{-1}\right)\)

- \( \mu|\beta, l_{\mu}, \eta_{\mu}, \Delta, \sigma^2, X, y \sim \text{MVN}\left(\begin{bmatrix} K^{-1}_{\mu} + (\sigma^2 I)^{-1} \end{bmatrix}^{-1} \begin{bmatrix} K^{-1}_{\mu} + (\sigma^2 I)^{-1} \end{bmatrix}^{-1} \right)^{-1} \begin{bmatrix} (\sigma^2)^{-1} (y - \Delta A) + K^{-1}_{\mu} X \beta \end{bmatrix}, \begin{bmatrix} K^{-1}_{\mu} + (\sigma^2 I)^{-1} \end{bmatrix}^{-1} \right)\)
To obtain the conditional posterior for $\Delta$, we first define some notation. Recall, $A$ is the vector of treatment indicators for all the subjects, and let $M$ denote a square matrix. $A^T \odot M$ indicates $A$ is multiplied element-wise to each column of $M$ while $M \odot A$ indicates $A$ is multiplied element-wise to each row of $M$.

$$
\Delta|\mu, l_\Delta, \eta_\Delta, \sigma^2, X, y \sim MVN \left( \left[ K^{-1}_\Delta + A^T \odot (\sigma^2 I)^{-1} \odot A \right]^{-1} A^T \odot (\sigma^2 I)^{-1} (y - \mu), \left[ K^{-1}_\Delta + A^T \odot (\sigma^2 I)^{-1} \odot A \right]^{-1} \right)
$$

The posterior distribution for the treatment effects of all subjects has covariance matrix $\left[ K^{-1}_\Delta + A^T \odot (\sigma^2 I)^{-1} \odot A \right]^{-1}$. Because it is difficult to write out the inverses, we obtain each element for the simple case with two subjects and a single covariate $X$. Let $A_1 = 1$ and $X_1$ denote the treatment status and covariate for subject 1 and $A_2 = 0$ and $X_2$ denote the treatment status and covariate for subject 2, so that there is a treated subject and a control subject. The covariance matrix is given by

$$
\left[ K^{-1}_\Delta + A^T \odot (\sigma^2 I)^{-1} \odot A \right]^{-1} = \begin{bmatrix}
\frac{\sigma^2 \eta^2_{11}}{\sigma^2 + \eta^2_1} & \sigma^2 \eta^2_{12} \exp \left\{ -\frac{1}{2} \left( \frac{X_1 - X_2}{l_\Delta} \right)^2 \right\} \\
\sigma^2 \eta^2_{21} \exp \left\{ -\frac{1}{2} \left( \frac{X_1 - X_2}{l_\Delta} \right)^2 \right\} & \frac{\sigma^2 \eta^2_{22}}{\sigma^2 + \eta^2_2} \end{bmatrix}
$$

Subject 2’s variance increases the further $X_2$ is from $X_1$ since a smaller amount would be subtracted from the second component of the product of $\left[ K^{-1}_\Delta + A^T \odot (\sigma^2 I) \right]^{-1}$, that is, $\left[ 1 - \frac{\eta^2_2}{\sigma^2 + \eta^2_2} \exp \left\{ -\left( \frac{X_1 - X_2}{l_\Delta} \right)^2 \right\} \right]$ becomes larger when $|X_1 - X_2|$ increases. Thus, there is greater variability in the treatment effect posterior as $|X_1 - X_2|$ increases. We note that to illustrate the key idea about uncertainty increasing the more dissimilar the covariates, we worked out the expression conditioning on $\mu(X)$. Therefore, for a treated subject (here, Subject 1), there is nothing to learn from control subjects, so the variability of the treated subject does not depend on the control subject’s data. On the other hand, for a control subject (Subject 2), information about $\Delta(X)$ must come from the other subject, so that the proximity of their covariate values, $X_1$ and $X_2$, will matter in terms of how certain we are about the estimate of $\Delta(X_2)$. Because outcome information from treated subjects is driving estimation of $\Delta$, we can expect more uncertainty when estimating this parameter for control subjects. For the off-diagonal elements, larger distances in the covariates result in smaller covariances so that there is less information that can be learned from the other person, which also corresponds to larger variability. This illustrates why with this prior specification we expect regions with little to no overlap to result in more uncertainty when it comes to causal effect estimation.

### 2.3.2 Metropolis within Gibbs algorithm for posterior inference

In this section, we briefly describe the steps of the algorithm for obtaining posterior sampling of the parameters of interest, $\mu$ and $\Delta$, and the hyperparameters. Details regarding the steps may be found in Appendix B in Data S1. For the hyperparameters in the GP priors ($l_\mu, \eta_\mu, l_\Delta$, and $\eta_\Delta$) and $\sigma^2$, we use a Metropolis-Hastings step for each one and update their values based on current values of the other parameters using an acceptance ratio. A candidate value is drawn from the proposal distribution—we employ a truncated normal distribution centered at the previous value with variance $\tau^2$, a tuning parameter, and bounded below at 0. We tune the variance parameters of the proposal distribution so that the jump sizes reflect spread in the posterior and the corresponding chain trace varies quickly around the mean.\(^{30,31}\) Convergence is assessed using time-series plots for each parameter to understand the number of MCMC iterations needed to observe stabilization of chains. The acceptance ratio compares the value of the posterior at the candidate value with that at the previous value. We randomly generate a value from the standard uniform distribution $U \sim \text{Unif}(0, 1)$; if $U$ is less than or equal to the ratio, the candidate value is accepted as the parameter value at the current iteration. Otherwise, the parameter value is set to its value from the previous iteration. For $\beta, \mu, \text{ and } \Delta$, at each iteration, their new value is drawn from their respective conditional distributions given current values of all the other parameters.

The chain is run until the number the posterior draws after thinning and burn-ins (say, $J$) is reached. The effect estimate at iteration $j$ is obtained as the mean over the elements of the $\Delta^{(j)}$ vector (ie, the average across all subjects): $\psi^{(j)} = \frac{1}{J} \sum_{j=1}^{J} \Delta^{(j)}$. The average treatment effect is then estimated as the mean of the posterior draws of $\psi$, $\Psi = \frac{1}{J} \sum_{j=1}^{J} \psi^{(j)}$, so that estimation for continuous outcomes may be obtained directly from the posterior distribution of $\Delta$. 
2.4 Extension to binary outcomes

In this section, we extend our model to dichotomous outcomes where \( Y \) may take on the values of either 0 or 1. Let \( \theta = \{ \mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta \} \). The probit model assigns to each \( X_i \) the variable \( Y_i \in \{0, 1\} \) using \( P(Y_i = 1) = \Phi(f(X_i, A_i, \theta)) = \Phi(\mu(X_i) + \Delta(X_i)A_i) \), where \( \Phi \) is the standard normal cumulative distribution function. Assuming the same priors for the parameters in \( \theta \) as those for the continuous outcome case, the posterior is

\[
p(\theta|X, A, Y) \propto p(Y|X, A, \theta) p(\theta) \propto \prod_{i=1}^{n} \Phi(f(X_i, A_i, \theta))^{Y_i} (1 - \Phi(f(X_i, A_i, \theta)))^{1-Y_i} p(\theta).
\]

Sampling \( \theta \) from this form is difficult. Thus, we consider the model augmented with a random variable \( Z_i \). Specifically, we define independent latent variables \( Z_i \), where each \( Z_i \) is normally distributed. Then the augmented probit model has the hierarchical structure as follows:

\[
Y_i = \begin{cases} 
1, & \text{if } Z_i > 0 \\
0, & \text{if } Z_i \leq 0 
\end{cases}
\]

\[
Z_i|\theta, X_i = \mu_i + \Delta_iX_i + e_i, e_i \sim N(0, 1) \]

\[
\theta \sim p(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta).
\]

Here, \( Y_i \) is deterministic conditional on the sign of \( Z_i \). Under the augmented model, \( P(Y_i = 1|X_i, A_i, \theta) = \Phi(f(X_i, A_i, \theta)) \), so the two models give the same inference. We will employ the augmented model in sampling of the parameters of interest and the latent variables. The joint posterior of latent variables \( Z \) and model parameters \( \theta \) given data \( X, A, Y \) is

\[
p(Z, \theta|X, A, Y) \propto p(Z|X, A, Y, \theta) p(\theta)
\]

\[
\propto p(Z|\mu, \Delta, A, X, Y) p(\mu|\beta, l_\mu, \eta_\mu, X) p(\beta)p(l_\mu)p(\eta_\mu)p(\Delta|l_\Delta, \eta_\Delta)p(l_\Delta)p(\eta_\Delta),
\]

where \( p(Z|\mu, \Delta, A, Y) = N(Z|\mu + \Delta A, 1) \left[ I(Y = 1)I(Z > 0) + I(Y = 0)I(Z \leq 0) \right] \).

Note that \( \theta \) is not dependent on \( Y \) given \( Z \), so the conditional posterior of the model parameters \( \theta \) is

\[
p(\theta|Z, X) \propto p(\theta)N(Z|\mu + \Delta A, 1).
\]

The conditional posterior of the latent variable \( Z_i \) is

\[
Z_i|\theta, Y_i, X_i \sim \left\{ \begin{array}{ll}
\text{TN(mean = } & \mu_i + \Delta_iX_i, \text{SD = 1, lower = 0, upper = } \infty), \\
\text{TN(mean = } & \mu_i + \Delta_iX_i, \text{SD = 1, lower = } \infty, \text{ upper = 0})
\end{array} \right., \quad \text{if } Y_i = 1
\]

\[
\left\{ \begin{array}{ll}
\text{TN(mean = } & \mu_i + \Delta_iX_i, \text{SD = 1, lower = 0, upper = } \infty), \\
\text{TN(mean = } & \mu_i + \Delta_iX_i, \text{SD = 1, lower = } \infty, \text{ upper = 0})
\end{array} \right., \quad \text{if } Y_i = 0
\]

Estimates of parameters are obtained by modifying the Metropolis within Gibbs algorithm such that \( Z_i \) takes the place of the continuous outcome and an additional step is used to sample \( Z_i \) from a truncated normal distribution.

Our interest is in the causal risk difference, \( \Psi = P(Y(1) = 1) - P(Y(0) = 1) \). The posterior for \( \Psi \) can be obtained as follows. From the Gibbs sampler, we will have stored \( J \) draws of \( \mu \) and \( \Delta \) (after discarding burn-ins and thinning). At each iteration \( j \), we obtain a draw of each potential outcome via computation. The probability of outcome under treatment, \( P_1^{(j)} = P(Y(1) = 1)^{(j)} \), is

\[
P_1^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mu_1^{(j)} + \Delta_1^{(j)}),
\]

and the probability of the outcome in the absence of treatment, \( P_0^{(j)} = P(Y(0) = 1)^{(j)} \), is

\[
P_0^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mu_0^{(j)}).
\]
Then the effect estimate at iteration \( j \) of the MCMC chain is \( \Psi^{(j)} = p_1^{(j)} - p_0^{(j)} \). The estimate of the risk difference is the average difference in proportions over the \( J \) posterior samples: \( \Psi = \frac{1}{J} \sum_j \Psi^{(j)} \).

### 3 | SIMULATION STUDIES

Simulation studies were used to assess the performance of the GP model for scenarios with varying degrees of nonoverlap. In these, we considered both linear and nonlinear response surfaces with the latter including treatment heterogeneity and interactions between covariates. We compared our GP approach to the following methods:

- **BCF** Bayesian causal forest with the prognostic and treatment components as functions of covariates and propensity scores, as proposed by Hahn et al.\(^{21}\)
- **BART-Stratified** Separate BART models are fit to treated and control subjects using covariates only and potential outcomes are estimated as the expected value of the function.\(^{35}\)
- **BART-Single** untrimmed BART as implemented by Nethery et al in which the treatment variable and propensity score are included as covariates, and potential outcomes are estimated with posterior predictive distributions.\(^{15,35}\) A single model is fit for the entire sample.
- **BART+SPL** The method proposed by Nethery et al for nonoverlap using the recommended \( a = 0.1, b = 10 \) to define the region of overlap based on propensity scores.\(^{15}\)
- **GLM** generalized linear model regression of outcome on main effects of treatment indicator and covariates with identity link for continuous outcomes and probit link for binary outcomes.

For the Bayesian methods, MCMC specifications include 10 000 burn-ins and 5000 iterations after burn-ins, in which every fifth is kept, yielding 1000 posterior draws of the average treatment effect.

We use 1000 replications for the simulations. For each simulated data set, we obtain 1000 posterior estimates (after discarding burn-ins and thinning) of treatment effect by averaging over all subjects. Specifically, for each replication \( k \), we have 1000 posterior draws of the treatment component (average over individual causal effects at each iteration). Denote the estimate of the treatment effect, the mean over the posterior draws, by \( \Psi_k \). The standard deviation \( SD_k \) and 95% credible intervals \( CI_k \) are obtained from these 1000 posterior draws. Over the 1000 replications we compute several quantities to measure performance:

\[
\text{ATE} = \frac{1}{1000} \sum_{k=1}^{1000} \Psi_k \\
\%	ext{Bias} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{\Psi_k - \text{ATE}_{\text{true}}}{\text{ATE}_{\text{true}}} \cdot 100 \\
\text{SE} = \sqrt{\frac{1}{1000-1} \sum_{k=1}^{1000} (\Psi_k - \text{ATE})^2} \\
\text{Coverage} = \frac{1}{1000} \sum_{k=1}^{1000} I(\text{ATE}_{\text{true}} \in CI_k)
\]

**Bias** = \( \frac{1}{1000} \sum_{k=1}^{1000} (\Psi_k - \text{ATE}_{\text{true}}) \)

**SD** = \( \frac{1}{1000} \sum_{k=1}^{1000} SD_k \)

**MSE** = \( \frac{1}{1000} \sum_{k=1}^{1000} (\Psi_k - \text{ATE}_{\text{true}})^2 \)

These simulations were conducted in R.\(^{36}\)

#### 3.1 | Continuous outcome

For the setting with a continuous outcome, we generate treatment indicator \( A \), two continuous covariates, and one binary covariate for \( n = 500 \) subjects. Our data generating model is specified as follows. First, treatment status is simulated as \( A \sim \text{Bernoulli}(0.5) \), and covariates are simulated based on treatment received. If \( A = 1 \), the covariates are generated as \( X_1 \sim N(\mu_1, 1), X_2 \sim N(\mu_2, 1), X_3 \sim \text{Bernoulli}(p) \); if \( A = 0 \), \( X_1 \sim N(0, 1), X_2 \sim N(2, 1), X_3 \sim \text{Bernoulli}(0.4) \). We consider two different outcome models.

1. \( Y_1 \sim N(1 - 2X_1 + X_2 - 1.2X_3 + 2A, 1) \).
2. \( Y_2 \sim N(-3 - 2.5X_1 + 2X_2^2A + \exp(1.4 - X_2A) + X_2X_3 - 1.2X_1 - 2X_2A + 2A, 1) \).

Different combinations of \( \mu_1, \mu_2, \) and \( p \) are chosen to control the amount of covariate nonoverlap in the sample. We consider settings with some nonoverlap \((\mu_1 = 1, \mu_2 = 2, P = .5)\) and substantial nonoverlap \((\mu_1 = 1, \mu_2 = 3, P = 0.6)\). In
the first model, the outcome is linearly related to covariates and treatment. In this case, for any combination of covariates, the treatment effect is the same—that is, treatment effect is homogeneous. We expect all methods to have decent performance since there are no interactions between covariates and treatment in the outcome model. Data generating model 2 incorporates nonlinearity and treatment heterogeneity. For instance, as \( X_1 \) values increase, the outcome \( Y \) for treated subjects tends to increase while \( Y \) values for control subjects tends to decrease. This leads to treatment effects that are larger in magnitude at larger \( X_1 \) values. Further, there is nonoverlap for these \( X_1 \) values since only treated subjects are observed in this region. The combination of treatment heterogeneity and nonoverlap makes it difficult to assess the treatment effect; parametric models would find it especially challenging to capture the true relationships.

We also consider the set of simulation scenarios in Nethery et al since our primary comparative method is BART+SPL.\(^{15}\) We use scenarios for which BART+SPL had the best performance; this method underperforms when propensity scores are misspecified as shown by simulation results from Nethery et al that considered misspecified propensity score models.\(^{15}\) Specifically, they show that larger bias and lower coverage resulted when estimated propensity scores from logistic regression are used. For \( n = 500 \) subjects, half are assigned treatment \( A = 1 \) and half are assigned to \( A = 0 \). Here, \( c \) controls the degree of overlap. The values of \( c \) that are considered are 0, 0.35, 0.7, where larger values correspond to greater extents of nonoverlap. Covariates are generated based on treatment assignment. For treated subjects \((A = 1)\), the covariates are generated with \( X_1 \sim Bernoulli(0.5), X_2 \sim N(2 + c, (1.25 + 0.1c)^2)\). For control subjects \((A = 0)\), \( X_1 \sim Bernoulli(0.4), X_2 \sim N(1, 1)\). The true propensity score is calculated based on density functions as follows:

\[
\text{True PS} = \frac{N(X_2; \mu = 2 + c, \sigma = 1.25 + 0.1c) \cdot \text{Ber}(X_1; p = 0.5)}{N(X_2; \mu = 2 + c, \sigma = 1.25 + 0.1c) \cdot \text{Ber}(X_1; p = 0.5) + N(X_2; \mu = 1, \sigma = 1) \cdot \text{Ber}(X_1; p = 0.4)}
\]

The true potential outcomes under control and under treatment for all subjects are generated:

\[
Y(0) = -1.5X_2,
\]
\[
Y(1) = \frac{-3}{1 + \exp(-10(X_2 - 1))} + 0.25X_1 - X_1X_2.
\]

Then the true treatment effect for each person is \( Y(1) - Y(0) \), so that there is a “true” ATE value for each simulated data set (say \( ATE_{true,k} \) for the \( k \)th replication).

\[
ATE_{true,k} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)
\]

The observed outcome is taken to be \( Y_i = A_iY_i(1) + (1 - A_i)Y_i(0) \).

Because there is a true ATE value for each replication indexed by \( k \), for \( k = 1, \ldots, 1000 \), we modify the bias and coverage metrics to accommodate different true values across the replications.

\[
\text{Bias} = \frac{1}{1000} \sum_{k=1}^{1000} (\Psi_k - ATE_{true,k})
\]
\[
\% \text{ Bias} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{\Psi_k - ATE_{true,k}}{|ATE_{true,k}|} \cdot 100
\]
\[
\text{Coverage} = \frac{1}{1000} \sum_{k=1}^{1000} I(ATE_{true,k} \in CI_k).
\]

3.1.1 Results

Simulation results for the nonoverlap scenarios in which the continuous outcomes are generated with a linear response surface are presented in Table 1. All methods provide estimates with low bias except BART-Stratified in the setting with some nonoverlap. We observe differences in the variability of the estimates. In particular, in the setting with some nonoverlap, the GP model’s estimate of variability is closest to that obtained from linear regression (the gold standard in this case). As the extent of nonoverlap increases, the variability obtained from the GP model increases to account for the
TABLE 1  Effect estimates for nonoverlap scenarios involving a linear response surface across methods

| Method      | ATE  | Bias | % Bias | SD       | SE    | MSE   | Coverage |
|-------------|------|------|--------|----------|-------|-------|----------|
| Some nonoverlap |      |      |        |          |       |       |          |
| GP          | 1.968| −0.032| −1.604| 0.102   | 0.100 | 0.011 | 0.948    |
| BCF         | 2.002| 0.002| 0.082  | 0.112   | 0.104 | 0.011 | 0.961    |
| BART-stratified | 1.904| −0.096| −4.820| 0.111   | 0.113 | 0.022 | 0.855    |
| BART-single | 2.011| 0.011| 0.561  | 0.115   | 0.106 | 0.011 | 0.968    |
| BART+SPL    | 2.015| 0.015| 0.729  | 0.156   | 0.118 | 0.014 | 0.980    |
| Linear model | 1.998| −0.002| −0.110| 0.101   | 0.098 | 0.010 | 0.945    |

Substantial nonoverlap

| Method      | ATE  | Bias | % Bias | SD       | SE    | MSE   | Coverage |
|-------------|------|------|--------|----------|-------|-------|----------|
| GP          | 1.971| −0.029| −1.458| 0.115   | 0.110 | 0.013 | 0.947    |
| BCF         | 1.971| −0.029| −1.428| 0.130   | 0.118 | 0.015 | 0.962    |
| BART-stratified | 1.946| −0.054| −2.684| 0.136   | 0.133 | 0.020 | 0.927    |
| BART-single | 1.985| −0.015| −0.772| 0.137   | 0.122 | 0.015 | 0.965    |
| BART+SPL    | 1.976| −0.024| −1.203| 0.288   | 0.175 | 0.031 | 0.997    |
| Linear model | 1.999| −0.001| −0.070| 0.112   | 0.108 | 0.012 | 0.955    |

Note: The true ATE is 2 for both degrees of nonoverlap.

Abbreviations: ATE, average treatment effect; BART, Bayesian additive regression trees; BCF, Bayesian causal forest.

greater uncertainty in those regions while maintaining nominal coverage. On the other hand, with increasing amounts of nonoverlap, the BART+SPL method results in doubled MSE and coverage very close to 1 (indicating overcoverage).

Simulation results for data simulated using the second data generating model are given in Appendix C in Data S1. For the two degrees of nonoverlap, the GP model gave percent biases of −10.612% and −27.153%, although these were the smallest biases among the comparator methods. For this complex data scenario, linear regression has the worst performance as expected. The other nonparametric models, BCF, the two BART models, and BART+SPL, all underestimate the average treatment effect with BART+SPL being the most biased. Furthermore, coverage was smaller than nominal for all the approaches. Note that the variability in the estimates of average treatment effect from the GP model again increase as the amount of nonoverlap increases as shown by SD values. Thus, while the GP model performed the best in terms of bias and efficiency, as reflected in its MSE metric, among the approaches considered under this data generating model, we note that there are certain settings with nonlinearities and interactions that may pose challenges to treatment effect estimation even for nonparametric models.

Table 2 presents the simulation results using the scenarios from Nethery et al. For each degree of nonoverlap, the GP model results in the smallest bias—even under the settings where BART+SPL was previously shown to perform best.15 The GP model has coverage that is nearly the same as that of BART+SPL but has variability estimates that are smaller. The high coverage for the GP model indicates that it both accurately estimates the truth and translates the uncertainty from nonoverlap regions into higher variability. The greater than nominal coverage from the GP model for these scenarios may be due to the absence of an error term when outcomes were generated. (We have considered the same simulation scenarios but with errors included when generating the potential outcomes. For the settings with c = 0, c = 0.35, and c = 0.70, the coverage estimates from the GP model were 0.978, 0.955, and 0.940, respectively, while the coverage estimates from BART+SPL were 0.990, 0.985, and 0.978, respectively. Thus, the randomness in the outcome generating model resulted in the GP model providing closer to nominal coverage.) In Table 2, we see that estimates provided by the GP model had smaller bias than estimates from BCF. The increase in the variability estimates as the amount of nonoverlap grows is larger for the GP model than for the BART-only models (BCF, BART-Stratified, and BART-Single), better reflecting the extent of nonoverlap. Given the nonlinearity and interactions specified in the outcome model, the parametric linear regression has the worst performance as expected.

3.1.2 Illustration of individual causal effect estimates

To illustrate individual causal effects obtained by each method, we use one simulated data for each nonoverlap scenario. For the linear response surface under substantial nonoverlap, we plot estimates of posterior mean and posterior SD of
individual causal effects across the iterations for each subject (Figure 1). The analogous figure for the moderate nonoverlap setting is included as Figure S1. The GP model provides estimates of individual causal effects that are close to 2 (the truth) indicating high precision, while those provided by BART-Single and BART+SPL are highly variable. Although the average treatment effect given by BART+SPL is close to the true value, the estimates of individual treatment effects range from −0.36 to 4.98, reflecting great variability.

The GP model tends to provide the lowest posterior SDs in regions of overlap. These estimates are only slightly greater than estimates of variability from linear regression, which would be correctly specified for this scenario. Because BART-Stratified models the treatment and control groups separately, baseline variability in estimates of treatment effect is higher. Notably, the Bayesian approaches show increases in variability in areas of increasing nonoverlap, indicating that they are capturing the greater uncertainty in those regions. The GP model’s estimates of individual-level SDs are greater than estimates of variability from linear regression, which would be correctly specified for this scenario.

Further, the few treatment subjects relative to the number of control subjects with PS near 0 is reflected in the larger variability as estimated by the GP model for these regions. BART+SPL does not take this into account for the specified $a$ and $b$ values.

---

**TABLE 2** Performance of the methods for nonoverlap scenarios from Nethery et al that employ true propensity scores

| Setting | Method      | ATE  | Bias | % Bias | SD    | SE    | MSE   | Coverage |
|---------|-------------|------|------|--------|-------|-------|-------|----------|
| $c = 0$ | GP          | −0.264 | 0.001 | 0.326  | 0.024 | 0.051 | $8.916 \times 10^{-5}$ | 1.000      |
|         | BCF         | −0.296 | −0.031 | −12.491 | 0.012 | 0.057 | 0.002 | 0.369      |
|         | BART-stratified | −0.304 | −0.039 | −15.643 | 0.014 | 0.056 | 0.002 | 0.378      |
|         | BART-single | −0.287 | −0.022 | −8.981  | 0.050 | 0.056 | 0.001 | 0.990      |
|         | BART+SPL    | −0.256 | 0.009 | 3.486   | 0.063 | 0.054 | $3.216 \times 10^{-4}$ | 1.000      |
|         | Linear model | −0.330 | −0.065 | −26.642 | 0.082 | 0.073 | 0.008 | 0.941      |
| $c = 0.35$ | GP          | −0.190 | −0.002 | −1.583  | 0.030 | 0.057 | $2.910 \times 10^{-4}$ | 0.998      |
|         | BCF         | −0.242 | −0.054 | −33.446 | 0.016 | 0.067 | 0.004 | 0.257      |
|         | BART-stratified | −0.263 | −0.075 | −45.849 | 0.020 | 0.067 | 0.007 | 0.204      |
|         | BART-single | −0.231 | −0.043 | −26.700 | 0.054 | 0.068 | 0.004 | 0.905      |
|         | BART+SPL    | −0.173 | 0.015 | 8.714   | 0.073 | 0.060 | $7.744 \times 10^{-4}$ | 1.000      |
|         | Linear model | −0.363 | −0.175 | −110.756 | 0.093 | 0.086 | 0.037 | 0.546      |
| $c = 0.70$ | GP          | −0.095 | −0.009 | −54.992 | 0.041 | 0.066 | $9.488 \times 10^{-4}$ | 0.995      |
|         | BCF         | −0.172 | −0.087 | −333.973 | 0.020 | 0.081 | 0.011 | 0.189      |
|         | BART-stratified | −0.215 | −0.130 | −472.294 | 0.026 | 0.083 | 0.021 | 0.123      |
|         | BART-single | −0.158 | −0.073 | −282.576 | 0.058 | 0.085 | 0.010 | 0.705      |
|         | BART+SPL    | −0.058 | 0.026 | 72.858  | 0.087 | 0.070 | 0.002 | 0.998      |
|         | Linear model | −0.402 | −0.316 | −1128.419 | 0.106 | 0.102 | 0.110 | 0.137      |

Abbreviations: ATE, average treatment effect; BART, Bayesian additive regression trees; BCF, Bayesian causal forest; GP, Gaussian process.
3.1.3 Sensitivity to specification of the GP prior

In the proposed GP model, prior distributions are specified for the hyperparameters instead of setting them to specific values in an effort to reduce sensitivity to these specifications. Further, a choice must be made about the covariance function $\mathcal{K}$. We explore the sensitivity of the model to these choices using two representative simulation scenarios—the same nonoverlap scenario with outcome generating model $Y_1$ and the nonoverlap setting from Nethery et al with $c = 0.35$. The first one involves a small degree of nonoverlap and a linear response surface while the second one involves a moderate degree of nonoverlap and a nonlinear response surface, reflecting varying data complexity.

For our primary implementation of the GP prior, we employed gamma (2,1) as the hyperprior for the $l$ and $\eta$ parameters—that is, we set $\alpha_\mu, \alpha_{\alpha_\mu}, \alpha_{\alpha_\eta}, \alpha_{\alpha_\eta} = 2$ and $\beta_{l_\mu}, \beta_{l_\eta}, \beta_{l_\eta}, \beta_{l_\eta} = 1$. These values corresponds to a distribution that is
right skewed and gives large probabilities to values near 0. To see if the model is sensitive to these hyperprior speci-
cfications, we examine the performance when \( l \) and \( \eta \) are assumed to have a gamma(4, 0.5) prior, such that \( \alpha_l, \alpha_\eta, \alpha_{l_\Delta}, \alpha_{\eta_\Delta} = 4 \) and \( \beta_l, \beta_\eta, \beta_{l_\Delta}, \beta_{\eta_\Delta} = 0.5 \). This distribution is centered at larger values and has a much larger spread (Figure S4). Simulation results comparing these two specifications are given in Table 1 in Appendix D in Data S1. Measures of bias, variability, and coverage are similar for these different hyperprior specifications.

Next, we compare different specifications of the covariance function for these two representative simulation settings. We compare the SQEXP covariance function to the rational quadratic, Matérn, and Ornstein-Uhlenbeck (exponential) covariance functions. For these specifications, the hyperparameters are given a gamma (2,1) prior and sampled using MCMC. Results are provided in Table 2 in Appendix D in Data S1. Again, estimates are in the same ballpark across the different covariance functions specifications, suggesting that the model is relatively insensitive to which kernel is used.
3.2 Binary outcomes

In simulation studies for binary outcomes, the treatment indicator \( A \) and covariates \( X_1, X_2, \) and \( X_3 \) are generated identically to the continuous case. We again consider two levels of nonoverlap and two outcome generating distributions. Outcomes are drawn from the Bernoulli distribution with proportion parameter that depends on covariate values and treatment received.

- \( Y_{1B} \sim \text{Bernoulli}(\Phi(-1 - 2X_1 + X_2 - 1.2X_3 + 2A)) \)
- \( Y_{2B} \sim \text{Bernoulli}(\Phi(-3 - 2.5X_1 + 2X_2^2 + \exp(1.4 - X_2A) + 1X_2X_3 - 1.2X_3 - 2X_3^2 + A)). \)

Detailed results of these simulations are provided in Appendix E in Data S1. In brief, the GP model gives similar bias and higher efficiency compared to the other methods and maintains its coverage when nonlinear and interaction terms are added to the outcome models.

3.3 High-dimensional covariate setting

Covariate nonoverlap is more likely to occur when there are many variables that are controlled for. To explore this setting, we simulate data according to the high dimensional scenarios employed in Nethery et al and compare the performance of the GP model to our primary comparator, the BART+SPL method.\(^{15} \) Implementation of the BART+SPL approach follows the original procedure used by Nethery et al when considering the high-dimensional setting. Again, half of the \( N = 500 \) subjects are assigned to treatment \( (A = 1) \) and the other half are placed in the control group \( (A = 0) \). 10 confounders (variables associated with both treatment and outcome) are generated. For treated subjects \( (A = 1) \), the covariates are generated with \( X_1, \ldots, X_5 \sim \text{Bernoulli}(0.45), X_6, \ldots, X_{10} \sim N(2, 4) \). For control subjects \( (A = 0) \), \( X_1, \ldots, X_3 \sim \text{Bernoulli}(0.4), X_6, \ldots, X_{10} \sim N(1.3, 1) \). The true potential outcomes under control and under treatment for all subjects are generated as follows.

\[
Y(0) = 0.5(X_1 + X_2 + X_3 + X_4 + X_5) + 15(1 + \exp(-8X_6 + 1))^{-1} + X_7 + X_8 + X_9 + X_{10} - 5, \\
Y(1) = X_1 + X_2 + X_3 + X_4 + X_5 - 0.5(X_6 + X_7 + X_8 + X_9 + X_{10})
\]

Three scenarios are explored via simulations.

1. HD 1: Only the 10 confounders are included in the model.
2. HD 2: The 10 confounders and 25 additional randomly generated variables (not truly related to the outcome variable) are included in the model.
3. HD 3: The 10 confounders and 50 additional variables are included in the model.

Measure of bias, variability, and coverage for the GP model and the BART+SPL method are presented in Table 3. Bias is slightly higher from the GP model for these particular simulation scenarios. However, the MSE from the GP model

| T A B L E 3 Comparisons of performance of the GP model to the BART+SPL method for high-dimensional covariate settings |
|-------------------------------------------------------------|
| \( \text{ATE} \) | \( \text{Bias} \) | \( \% \text{Bias} \) | \( \text{SD} \) | \( \text{SE} \) | \( \text{MSE} \) | \( \text{Coverage} \) |
|-----------------|-------------|-------------|--------|--------|--------|--------------|
| HD 1 GP         | -16.688     | 0.732       | 4.203  | 0.353  | 0.513  | 0.684        | 0.430        |
| GP BART+SPL     | -17.631     | -0.212      | -1.218 | 0.667  | 1.877  | 3.462        | 0.524        |
| HD 2 GP         | -18.315     | -0.912      | -5.249 | 0.272  | 0.380  | 0.896        | 0.072        |
| GP BART+SPL     | -16.991     | 0.412       | 2.371  | 0.537  | 1.712  | 2.961        | 0.492        |
| HD 3 GP         | -18.125     | -0.715      | -4.113 | 0.300  | 0.405  | 0.588        | 0.354        |
| GP BART+SPL     | -16.856     | 0.554       | 3.169  | 0.513  | 1.548  | 2.671        | 0.486        |

Abbreviations: ATE, average treatment effect; BART, Bayesian additive regression trees; GP, Gaussian process.
is less than that from the BART+SPL method. While the bias and coverage from the BART+SPL method tends to get worse as additional variables are included in the model, there is no clear pattern for the GP model as the coverage is very low for the HD 2 scenario yet appears to bounce back up in the HD 3 scenario. Further investigation of the GP model’s performance and modification to its form or prior specifications may be needed to better accommodate high-dimensional covariate settings.

4 | APPLICATION TO STUDY OF RIGHT HEART CATHETERIZATION

We applied our GP approach to data on critically ill patients in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). Details on the study population and data collection have been previously described in Connors et al. We note that the purpose of this data example is to demonstrate our method and not to make clinical claims. These data are publicly available which will allow readers to readily replicate our results. In our analysis, we assess the effects of right heart catheterization (RHC) in the first 24 hours upon entry into study on survival for female subjects. The binary outcome of interest in this study is defined as

\[ Y_i = \begin{cases} 
1, & \text{if subject } i \text{ died within 180 days} \\
0, & \text{otherwise}
\end{cases} \]

The confounding variables of interest include age, race, years of education, income, medical insurance, primary disease category, Activities of Daily Living score, Duke Activity Status Index, do-not-resuscitate status, cancer status, SUPPORT model estimate of the probability of surviving 2 months, APACHE II score, coma score based on Glasgow on day 1, physiological measurements, and categories of comorbid illness. Of the 617 female subjects of interest, 137 received an RHC while 480 did not. In the treatment group, 62 died within 180 days, compared with 219 in the control group. Characteristics of the sample for analysis are provided in Appendix F in Data S1. We see that the RHC group tends to be younger on average, have higher income, are less likely to sign a do-not-resuscitate form, and have lower respiratory rates and lower PaCO2 on Day 1. Further, the proportions of people with pulmonary disease were significantly different between the RHC and non-RHC groups. Propensity scores were estimated using a BART probit model with treatment status as the response and all confounding variables as predictors. Figure 3 demonstrates nonoverlap in the tails.

![Histograms of estimated propensity scores for patients who received an right heart catheterization and those that did not](image-url)

**FIGURE 3** Histograms of estimated propensity scores for patients who received an right heart catheterization and those that did not
TABLE 4  Estimated average treatment effect of receiving the right heart catheterization

|          | ATE   | SE    | 95% CrI          |
|----------|-------|-------|------------------|
| GP       | 0.022 | 0.032 | [−0.032, 0.100]  |
| BCF      | 0.033 | 0.044 | [−0.047, 0.126]  |
| BART-Stratified | 0.035 | 0.049 | [−0.059, 0.124]  |
| BART-Single | 0.031 | 0.042 | [−0.049, 0.114]  |
| BART+SPL | 0.030 | 0.077 | [−0.111, 0.177]  |

Abbreviations: ATE, average treatment effect; BART, Bayesian additive regression trees; GP, Gaussian process.

To fit the Gaussian process model, we employed four chains with different initial values to estimate the parameters of interest. Specifically, each chain involved 10,000 burn-ins and 20,000 iterations after burn-ins, from which every 80th was kept in order to minimize autocorrelation. Combining these iterations, we obtained 1000 posterior draws for the parameters. We calculated a risk difference defined as the mean difference in probabilities of dying within 180 days from the start of study entry had the RHC been given vs had the RHC not been given, respectively. The risk difference estimate was 0.024 with 95% credible interval [−0.031, 0.098]. These estimates indicate that the 180-day survival of subjects who received the RHC did not differ significantly from that of patients who did not get RHC. Point estimates obtained from the comparator methods were found to be similar as shown in Table 4. The GP model resulted in narrower credible intervals, which is consistent with what we found in some of the simulation studies (Tables 1 and 2).

5 | DISCUSSION

In this paper, we develop a model that employs GP priors to address practical violations of the positivity assumption when estimating causal effects from observational data. Unlike matching or trimming approaches, our method allows inference about the original target population. Further, unlike previous extrapolation methods, our approach does not require specifying arbitrary cut-offs in order to define nonoverlap regions. Importantly, our GP approach better reflects the greater uncertainty around estimated causal effects that is expected in areas of less covariate overlap.

For complex outcome models containing nonlinearities and interactions, the GP model provided average treatment effect estimates with good performance in certain scenarios. For other settings that may be particularly challenging for valid estimation, the GP model resulted in bias but was shown to have better performance in comparison to the alternative models considered. This result may be attributed to the nonparametric nature of the GP model and the centering of the GP prior of the prognostic component on a linear model. By pulling the prior in the direction of the data, more accurate estimates were obtained in data sparse areas. Further, the form of our model places a direct prior on the treatment effect, which may be beneficial for incorporating prior knowledge of the treatment effect and for subsequent interpretation—the form of the posterior for treatment effects is known. The GP model also provided more accurate and precise estimates of individual causal effects, which were most likely due to its accommodation of each subject’s actual covariate values. For instance, for subjects in nonoverlap regions, we observed the GP model to be superior to the BART-only models in providing individual-level estimates that are close to the truth.

We emphasize the point that the type of positivity violation and the choice of approach that would be appropriate for addressing those violations depends on the clinical or study question of interest and the population of interest. If the objective is to account for all patient subgroups, that is, the set eligible for a particular treatment or intervention, then the population-level estimand is needed. There may be an extent of nonoverlap past which it would not be reasonable to obtain a population-level estimand—this may be a case when there are issues with study design or sampling, which should be dealt with before data are analyzed. Prior to fitting the GP model, standard exploratory analyses to assess covariate or propensity score overlap should be carried out and different definitions of overlap that have been proposed could be considered.38 High degrees of nonoverlap may indicate that the current differences between the groups may result in invalid comparisons and that an alternative target population may be of more scientific or clinical interest.
In their invited discussion to Hahn et al, Papadogeorgou and Li suggest the Gaussian process model may better address regions of poor overlap, where the GP model was shown to have larger measures of uncertainty. In their example, they consider a single covariate and then fit separate outcome models for the treatment and control groups where the functions for each are given a GP prior. In our proposed approach, instead of placing priors on the outcome models for treated subjects and control subjects, respectively, we propose utilizing two GP priors in the same model for the prognostic and treatment effect components. With this choice of modeling, we allow data from both the treatment and control groups to influence the model fitting and thus more information is utilized in estimating the treatment effects. Our study furthers their illustrations by exploring the performance of the GP prior to address covariate nonoverlap in more complex data scenarios involving different degrees of nonoverlap and varying numbers of covariates. Further, by employing hyperpriors for the hyperparameters in the GP prior in the implementation of our model, results may be less sensitive to the particular prior specification.

One current limitation to our approach is the potential lack of scalability to very large studies due to computational challenges. The computational complexity is $\mathcal{O}(n^3)$ due to inversion of matrices that have dimensions equal to the sample size. Further, the long run time may be due to the number of parameters being sampled in our Markov Chain Monte Carlo; our algorithm estimates hyperparameters rather than fixing them at constant values. In the covariance function, the same length-scale parameter is used for all the covariates. Modifying the kernel in the GP prior to allow for different length-scale parameters for different covariates may be beneficial especially when there is a large number of covariates. This may better capture the correlation between each covariate and outcome to reflect the relative importance of each variable and allow prior information regarding relevant confounders to be utilized. Allowing different length-scale parameters may improve performance of the model when there is a very large number of covariates. For instance, in sampling the length-scale parameters, the strength the effect of each covariate and outcome based on the data may be used to reduce the effect of less important variables; that is, variables that are less relevant for predicting the parameter will tend to have larger length-scales. This may be done in conjunction with variable selection and/or sparsity priors to address high-dimensional covariates, and such existing approaches may be adapted to our causal setting. Moreover, here we have focused on continuous and binary outcomes. We are currently considering extensions of the GP approach to other outcomes that are common in clinical studies such as censored survival outcomes and longitudinal outcomes. In the longitudinal setting, addressing positivity violations may pose particular challenges due to time-dependent confounding.

ACKNOWLEDGMENTS
The authors received no specific funding for this work.

CONFLICT OF INTEREST
The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT
The right heart catheterization dataset used in this study is openly available at http://hbiostat.org/data courtesy of the Vanderbilt University Department of Biostatistics. R functions to implement our proposed Gaussian process model are included.

ORCID
Angela Yaqian Zhu https://orcid.org/0000-0001-5187-2667

REFERENCES
1. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health. 2006;60(7):578-586. doi:10.1136/jech.2004.029496
2. Westreich D, Cole SR. Invited commentary: positivity in practice. Am J Epidemiol. 2010;171(6):674-677; discussion 678-681. doi:10.1093/aje/kwp436
3. Imbens GW, Rubin DB. Assessing Overlap in Covariate Distributions. Cambridge: Cambridge University Press; 2015: 309-336.
4. D’Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. J Econom. 2020;221(2):644-654. doi:10.1016/j.jeconom.2019.10.014
5. King G, Zeng L. The dangers of extreme counterfactuals. Polit Anal. 2006;14(2):131-159. doi:10.1093/pan/mpj004
40. Linkletter C, Bingham D, Hengartner N, Higdon D, Ye KQ. Variable selection for Gaussian process models in computer experiments. Technometrics. 2006;48(4):478-490.

**SUPPORTING INFORMATION**
Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Zhu AY, Mitra N, Roy J. Addressing positivity violations in causal effect estimation using Gaussian process priors. Statistics in Medicine. 2023;42(1):33-51. doi: 10.1002/sim.9600