A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation

Arman Oganisian1 | Nandita Mitra1 | Jason A. Roy2

1Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania
2Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, New Jersey

Correspondence:
Arman Oganisian, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA 19104. Email: aoganisi@upenn.edu

Funding information
American Cancer Society, Grant/Award Number: 124268-IRG-78-002-35-IRG; National Institute of General Medical Sciences, Grant/Award Number: R01GM112327

Abstract
Researchers are often interested in predicting outcomes, detecting distinct subgroups of their data, or estimating causal treatment effects. Pathological data distributions that exhibit skewness and zero-inflation complicate these tasks—requiring highly flexible, data-adaptive modeling. In this paper, we present a multipurpose Bayesian nonparametric model for continuous, zero-inflated outcomes that simultaneously predicts structural zeros, captures skewness, and clusters patients with similar joint data distributions. The flexibility of our approach yields predictions that capture the joint data distribution better than commonly used zero-inflated methods. Moreover, we demonstrate that our model can be coherently incorporated into a standardization procedure for computing causal effect estimates that are robust to such data pathologies. Uncertainty at all levels of this model flow through to the causal effect estimates of interest—allowing easy point estimation, interval estimation, and posterior predictive checks verifying positivity, a required causal identification assumption. Our simulation results show point estimates to have low bias and interval estimates to have close to nominal coverage under complicated data settings. Under simpler settings, these results hold while incurring lower efficiency loss than comparator methods. We use our proposed method to analyze zero-inflated inpatient medical costs among endometrial cancer patients receiving either chemotherapy or radiation therapy in the SEER-Medicare database.

KEYWORDS
Bayesian, causal inference, clustering, Dirichlet process, healthcare costs, nonparametrics, zero inflation

1 INTRODUCTION

Researchers across many fields are often interested in outcome prediction, clustering analysis, and causal inference. For example, researchers in personalized medicine are broadly concerned with forming out-of-sample outcome predictions given a subject’s covariates. Health economists are often interested in subgroup identification for resource allocation purposes and may turn to algorithms such as K-means. Policy researchers, on the other hand, focus on causality—estimating the average difference in outcomes that would have occurred under hypothetical policy interventions. All of these tasks become challenging in the presence of zero-inflated outcomes, multimodality, and extreme skewness. Structural zeros often need to be modeled: If causal treatment effect estimation is the goal, failing to capture a difference in prevalence of zeros between treatment groups may bias effect estimates. For prediction purposes, it is necessary to capture outcomes at the skewed high end of the distribution as well as predicting the structural zeros at the low end. Failing to do so would tarnish predictions at both tails. For clustering analyses, having to prespecify the number of
clusters—typically an unknown quantity—poses a significant challenge.

In this paper, we develop a Bayesian nonparametric (BNP) generative model that simultaneously predicts structural zeros as a function of covariates, captures skewness in both the outcome and continuous covariates, and induces a grouping of subjects into clusters with similar joint data distributions. The result is a flexible, multipurpose model that is broadly applicable to the tasks described above. We demonstrate the ability of our model to produce robust causal effect estimates via standardization—a common method for computing marginal causal contrasts while adjusting for measured confounders. This fully Bayesian approach allows uncertainty to propagate through to the causal estimates, allowing point and interval estimation of various causal contrasts such as mean differences and quantile causal effects. Moreover, posterior predictive checks around positivity—a key causal identification assumption—can be readily conducted using the model output.

In particular, we propose a Dirichlet process (DP) mixture of zero-inflated regressions. Each zero-inflated regression is a two-part model: a model for the probability of the outcome being zero and a regression for the continuous, nonzero outcomes. DP mixtures (Ferguson, 1973) are a class of BNP models that partition a complex joint distribution of the outcome and covariates into more homogeneous clusters. In our case, the cluster-specific conditional means are modeled using a zero-inflated regression. Unlike finite mixtures, DP mixtures assume there are infinitely many clusters in the population—removing the need to specify the number of clusters in advance. As many clusters are introduced as are needed to accommodate the complexity of the data. If the data are not complex and can be adequately fit with a parametric model, new clusters form less often. In this sense, our model is data adaptive—growing in proportion to the complexity of the data.

The flexibility and relative ease of constructing point and interval estimates for various types of contrasts are perhaps some of the reasons that BNP methods have been growing in popularity within the causal inference literature. For example, Bayesian additive regression trees (BART) (Chipman et al., 2010; Hill, 2011) have been used to estimate causal treatment effects. Dependent DP methods have been developed for estimating marginal structural models (Roy et al., 2014) and dynamic treatment regime models (Xu et al., 2016). DP mixture approaches for mediation analysis (Kim et al., 2017) and enriched Dirichlet process (Wade et al., 2014) mixture approaches to standardization have also been developed (Roy et al., 2018). However, these methods do not address the complications of zero-inflation discussed. We advance existing methodology by developing a BNP standardization approach that accounts for zero inflation.

Several factors distinguish our approach from the existing zero-inflated models outside of the causal inference literature. As opposed to the parametric Bayesian approach of Ghosh et al. (2006), our method is nonparametric and, therefore, better suited for complex data. Barcella et al. (2016) develop a DP mixture of Poisson regressions. Similarly, Kurz et al. (2019) apply a finite mixture of negative binomial regressions to utilization count data. Though these provide a flexible fit to count data, they are inappropriate for semi-continuous data. Linero et al. (2018) develop a semiparametric Bayesian model for semicontinuous outcomes. They use a two-part model—a probit model for the probability of a zero and a parametric density for nonzero outcomes. The mean functions of both models are jointly estimated using a BART-based model. In contrast, our model is fully nonparametric, DP based as opposed to BART based, and generative as opposed to conditional. That is, our model estimates the full joint data distribution rather than solely a conditional outcome distribution. The strength of DP-based procedures over BART-based procedures is that the former induces clustering—allowing us to capture multimodalities. Using generative models as opposed to conditional models provides a framework for flexibly imputing missing data, as was demonstrated by Roy et al. (2018).

Though broadly applicable, we motivate our approach throughout the paper by the analysis of medical cost outcomes—an important use case of our method. Zero inflation is the norm in cost data as patients may tend to have zero costs through mechanisms that depend on measured covariates and the assigned treatment. Medical costs also tend to be skewed by especially high-cost patients. Moreover, the joint distribution tends to be multimodal with groups of patients that exhibit different cost-covariate relationships. Legislators and regulators often make use of economic analyses comparing costs associated with proposed policy interventions. These comparisons are causal in nature and require robust statistical modeling while adjusting for confounders.

## 2 | DIRICHLET PROCESS MIXTURE OF ZERO-INFLATED REGRESSIONS

### 2.1 | A generative model

Consider observing data \( D = (D_i)_{i=1:n} = (Y_i, A_i, L_i)_{i=1:n} \) from \( n \) independently sampled subjects. The \( q \times 1 \) covariate vector, \( L_i \), contains both categorical and continuous covariates measured pretreatment. The scalar \( A_i \in \{0, 1\} \) denotes binary treatment assignment. The scalar outcome is \( Y_i \)—whose empirical distribution may exhibit excess zeros, skewness, and multimodality. We first define covariate vectors for subject \( i \) as \( x_i = (1, A_i, L_i)' \) and \( m_i = (1, L_i)' \). We specify a generative model—that is, a
model for the full joint \( p(D_i \mid \omega_i) = p(Y_i \mid A_i, L_i, \omega_i)p(A_i \mid L_i, \omega_i)p(L_i \mid \omega_i) \). Hierarchically this is given by,

\[
Y_i \mid A_i, L_i, \beta_i, \gamma_i, \phi_i \sim \pi(x'_i \gamma_i) \delta_0(y_i) + (1 - \pi(x'_i \gamma_i)) \cdot N(y_i \mid x'_i \beta_i, \phi_i)
\]

\[
A_i \mid L_i, \eta_i \sim \text{Ber} (\expit(m'_i \eta_i))
\]

\[
L_i \mid \theta_i \sim p(l_i \mid \theta_i)
\]

\[
\omega_i \mid G \sim G
\]

\[
G \mid \alpha, G_0 \sim DP(aG_0).
\]

Above, we define \( \omega_i = (\beta_i, \phi_i, \gamma_i, \eta_i, \theta_i) \) for compactness. The conditional distribution of the outcome, \( Y_i \), is modeled as a two-part mixture of a point-mass at 0, \( \delta_0(y_i) = I(y_i = 0) \), and a Gaussian distribution with mean \( x'_i \beta_i \) and variance \( \phi_i \).

This allows for a positive probability of the outcome being zero, \( P(Y_i = 0) = \pi(x'_i \gamma_i) = \expit(x'_i \gamma_i) \). This probability is modeled as a function of treatment and confounders using a logistic regression with a \( (q + 2) \times 1 \) parameter vector \( \gamma_i \). Separately, the conditional mean of nonzero outcomes is modeled using a regression with a \( (q + 2) \times 1 \) parameter vector \( \beta_i \).

Though we use logit links to model \( \pi(x'_i \gamma_i) \) and \( P(A_i = 1 \mid L_i, \eta_i) \), other links, such as the probit, could be used as well. In anticipation of subsequent application to causal estimation, we model treatment probability (ie, the propensity score) as a function of confounders, \( L_i \), using a logistic regression with a \( (q + 1) \times 1 \) parameter vector \( \beta_i \).

In application areas where outcomes are nonnegative and observed data are close to zero, a local Gaussian distribution that ignores the nonnegative nature of the outcome is undesirable. In these instances, we can proceed with the model as presented after log-transforming nonzero values—essentially assuming a log-normal distribution for these values. Web Appendix F in the supplement provides more details along with a proof-of-concept simulation.

### 2.2 Posterior sampling and hyperparameters

Using the Polya Urn (Blackwell and MacQueen, 1973) representation of the DP, it can be shown that the conditional posterior of \( \omega_i \) is given by (Muller et al., 2015)

\[
p(\omega_i \mid \omega_{1:(i-1)}, D) \propto \frac{1}{\alpha + i - 1} \left[ a p(D_i \mid \omega_i) G_0(\omega_i) + \sum_{j < i} p(D_j \mid \omega_j) \delta_{\omega_j}(\omega_i) \right],
\]

where \( D_i = (Y_i, A_i, L_i) \) is the data vector for the \( i \)th subject. The posterior clustering of patients is evident in Equation (2). Subject \( i \)'s parameter, \( \omega_i \), can equal one of the previously drawn parameters, \( \omega_j \), with probability proportional to the subject’s likelihood evaluation under \( \omega_j \): \( \sum_{j < i} p(D_j \mid \omega_j) \cdot \delta_{\omega_j}(\omega_i) \). Or, with probability proportional to \( a p(D_i \mid \omega_i) \), \( \omega_i \) can be a new, previously unseen parameter drawn from the prior \( G_0(\omega_i) \).

If subject \( i \) is quite unique so that its likelihood evaluation is low under the \( i - 1 \) existing parameters, then it is relatively more likely for this subject to be assigned its own set of parameters from the prior. Finally, note that as \( n \) gets large and \( i \) approaches \( n \), the prior probability \( a / (\alpha + i - 1) \) of the \( i \)th subject being assigned to a new cluster goes to zero. This property helps prevent overfitting.

The conditional posterior in Equation (2) forms the basis of a Metropolis-in-Gibbs sampler we use to sample \( \omega_{1:n} \) from the full posterior, \( p(\omega_{1:n} \mid D) = p(\omega_1 \mid D) \prod_{i=2}^{n} p(\omega_i \mid \omega_{1:(i-1)}, D) \). The sampler proceeds in the spirit of Neal’s Algorithm 8 (Neal, 2000) by introducing latent cluster membership indicators, \( c_{1:n} = (c_1, c_2, \ldots, c_n)' \).
for the subjects. We initialize the algorithm by partitioning subjects to one of \( K \) initial clusters. Each iteration \( t \), with \( K(t) \) occupied clusters indexed by \( k \), has two steps. First, conditional on \( c_i^{(t)} \), we draw from the posterior of each model parameter based on the likelihood contributions of all subjects with \( c_i^{(t)} = k \). Conditional on these updated parameter, \( \omega_i^{t+1} \), we update each assignment indicator

\[
c_i^{(t+1)} \mid c_i^{(t)}_{1:(t-1)} \sim \text{Cat}\left(\frac{1}{\alpha + i - 1} p(D_i|\omega_i^{t+1}), \ldots, \frac{1}{\alpha + n - 1} p(D_i|\omega_i^{0})\right).
\]

Above, \( \text{Cat}(\cdot) \) denotes the categorical distribution and \( \omega_0^{t+1} \sim G_0 \) is a draw from the prior taken at each iteration. Notice that in each iteration subject \( i \) has a \( \frac{\alpha}{\alpha + i - 1} p(D_i|\omega_i^{t+1}) \) probability of being assigned to a new cluster.

The two hyperparameters of the model in Equation (1) are the choice of base distribution, \( G_0 \), and the concentration parameters \( \alpha \). A requirement for the base distribution is that it be over the space of the parameters \( \omega_i = (\beta_i, \phi_i, \gamma_i, \eta_i, \theta_i) \). Prior independence is often assumed so that \( G_0 \) can be constructed as the product over parameter-specific priors. Conjugate priors for each parameter may be used, if possible, to simplify computation. The concentration parameter \( \alpha \) governs how frequently new clusters appear. It is often described as a prior sample size for a new cluster. Following previous analyses (Roy et al., 2018), we place a \( \alpha \sim \text{Gamma}(1, 1) \) prior on \( \alpha \) rather than set it at a particular value. Examples of specifying \( G_0 \) are given in Web Appendices D, E, and F for simulations, data analysis, and log-transform extension, respectively.

### 2.3 Posterior mode clustering in the presence of label switching

Often we may like to cluster patients using the posterior mode—allowing us to identify and summarize distinct groups in terms of observed characteristics. In mixture models, posterior mode inference on cluster assignment is complicated by label switching (Rodriguez and Walker, 2014)—the fact that cluster labels \( c_{1:n} \) do not have consistent meanings across Gibbs iterations. For example, at iteration \( t \), a new cluster, labeled cluster 2, may be proposed and all subjects previously in, say, cluster 1 may be reassigned to this new cluster. Even though the cluster label has changed from 1 to 2, the cluster still contains the same subjects. Therefore, naively using the mode of the \( T \) cluster indicators, \( c_i^{(1)}, \ldots, c_i^{(T)} \), as each subject as the mode assignment is problematic. To meaningfully cluster subjects based on posterior mode, we perform a deterministic relabeling of cluster indicators after posterior sampling (Dahl, 2006; Stephens, 2000). We compute for each iteration \( t \) an \( n \times n \) adjacency matrix with a one in the \((i, j)\)th entry, indicating patients \( i \) and \( j \) were clustered together, and zero, indicating otherwise. The elementwise mean of this matrix across the \( t \) iterations gives us a posterior mode matrix with \((i, j)\)th entry being the posterior probability of patients \( i \) and \( j \) being clustered together. To obtain cluster assignments, we select the adjacency matrix that is closest in the \( L_2 \) sense to the posterior mode matrix. More details regarding the relabeling is provided in Web Appendix B in the supplement.

### 3 Counterfactual prediction and estimating causal contrasts

#### 3.1 Review of counterfactuals and causal estimation

We first provide a motivating review of causal estimation before discussing our BNP standardization procedure. Consider observing \( D = (Y, A, L)_{i=1:n} \) as defined in Section 2.1 from some target population we wish to make inference about. Using potential outcome notation, let the random variable \( Y^A=a \) represent the potential outcome under treatment \( A = a \). The marginal causal effect of treatment on the outcome, \( \Psi = E[Y^A=1 - Y^A=0] \), can be computed via the method of standardization under the standard causal identification assumptions (Rubin, 1978) of ignorability, consistency, no interference, and positivity. Briefly, and in order, these assumptions require that all confounders are controlled for, that there is only one form of each treatment, that each patient’s outcome is independent of others’ treatment assignments, and that treatment assignment is not deterministic for any individual in the population. We provide a formal statement in Web Appendix C in the supplement.

In the Bayesian framework, standardization is conducted using the posterior predictive distribution of the outcome (Keil et al., 2017). Throughout, we use tildes to denote posterior predictive draws. Let \( \tilde{Y}^a \) denote the posterior predictive outcome under intervention \( A = a \) with predictive distribution \( p(\tilde{Y}^a|D) \). Also, let \( \tilde{L} \) denote a posterior predictive draw of confounders. If the causal assumptions hold, standardization under intervention \( A = a \) is given by

\[
E(\tilde{Y}^a|D) = \int_\theta \int_\beta \int_L E(\tilde{Y}^a|A = a, \tilde{L}, \beta)p(L|\theta)p(\beta, \theta|D) \, dL \, d\beta \, d\theta.
\]

\[ (3) \]

Above, \( \beta \) and \( \theta \) are parameter vectors that govern the conditional distribution of the outcome and the distribution of the confounders, respectively. This slightly differs from frequentist standardization by additionally averaging a prediction model for the outcome, \( E(\tilde{Y}^a|A = a, \tilde{L}, \beta) \), over
the posterior distribution of the parameters, \( p(\beta, \theta | D) \). Given \( T \) draws from the posterior \( (\theta^{(t)}, \beta^{(t)})_{t=1:T} \), we can compute Equation (3) by first drawing \( \tilde{Y}^{(t)} \sim p(\tilde{Y} | \theta^{(t)}) \), then computing \( E[Y | A = a, \tilde{L} = \tilde{L}^{(t)}, \beta^{(t)}] \). This yields a posterior distribution for the difference \( \{ \delta^{(t)} \}_{t=1:T} = \{ E[Y | A = 1, \tilde{L} = \tilde{L}^{(t)}, \beta^{(t)}] - E[Y | A = 0, \tilde{L} = \tilde{L}^{(t)}] \}_{t=1:T} \). The posterior mean \( \Psi = T^{-1} \sum_t \delta^{(t)} \) can be taken as a point estimate of \( \Psi \), while percentiles can be used for interval estimation. Standardization in Equation (3) crucially requires both a correctly specified regression for the outcome as well as an accurate estimate of the marginal confounder distribution. As correct specification is unlikely, robust estimation requires nonparametric modeling. This is especially the case in medical cost data—where multimodality, zero inflation, and skewness are unlikely to be captured by simple parametric models.

### 3.2 Sampling from the posterior predictive distribution

The model outlined in Equation (1) yields a flexible predictive distribution, which in turn yields robust causal effect estimates. Under standard causal identification assumptions, the posterior predictive distribution of potential outcome \( \tilde{Y}^a \) is given by

\[
p(\tilde{Y}^a | D) = \frac{a}{\alpha + n} \int_\omega \int_\tilde{L} p(\tilde{Y} | A = a, \tilde{L}, \omega) dP(\tilde{L} | \omega) dG_0(\omega)
\]

\[
+ \frac{1}{\alpha + n} \int_\omega \left[ \int_{\tilde{L}} p(\tilde{Y} | A = a, \tilde{L}, \omega_i) dP(\tilde{L} | \omega_i) \right] dP(\omega_1 | D).
\]

A derivation is provided in Web Appendix A in the supplement. Note that the particular forms of \( p(\tilde{Y} | A, \tilde{L}, \omega) \) and \( p(\tilde{L} | \omega) \) are specified in Equation (1). We can draw from this distribution via Monte Carlo. For each of the \( T \) posterior draws, \( (\omega^{(t)}_{1:n})_{t=1:T} \) from \( p(\omega_1 | D) \), draw from the conditional distribution \( l \sim p(\tilde{L} | \omega^{(t)}) \). Then, under intervention \( A = a \), we can draw from \( \tilde{Y}^{(t)}_a \sim p(\tilde{Y} | A = a, \tilde{L} = l, \omega^{(t)}_i) \). For the \( t \)th draw, the inner integral over \( \omega \) in Equation (4) can be evaluated numerically by drawing from the prior \( \omega_0 \sim G_0 \), then drawing confounders conditional on this prior draw \( l \sim p(l | \omega_0) \). This procedure yields predictive draws \( \{ \tilde{y}^{(t)}_a \}_{t=1:T} \).

After obtaining draws \( \{ \tilde{y}^{(t)}_a \}_{t=1:T} \), we can compute \( T \) draws of the difference \( \{ \delta^{(t)}_i \}_{t=1:T} = \{ \tilde{y}^{(t)}_a - \tilde{y}^{(t)}_0 \} \). Then, a BNP point estimate of \( \Psi \) is given by \( \Psi_{\text{BNP}} = E[\tilde{Y} | D] - E[\tilde{Y}^0 | D] \approx \frac{1}{T} \sum_t \delta^{(t)}_i \). Intervals can be constructed using percentiles of \( \{ \delta^{(t)}_i \}_{t=1:T} \). Quantile causal effects and counterfactuals (Xu et al., 2018) may also be computed from Equation (4). We estimate the posterior predictive cumulative distribution function (CDF) of the potential outcome under intervention \( a \) using the posterior predictive draws, \( F_a(v) = P(\tilde{Y}^a \leq v | D) \approx \frac{1}{T} \sum_{t=1}^T I(\tilde{y}^{(t)}_a \leq v) \). The inverse of this estimated CDF can be used to estimate quantile causal effects. For instance, the median causal effect can be estimated as the difference in median cost under two interventions \( F^{-1}_a(.5) - F^{-1}_0(.5) \). This contrast may be preferable to \( \Psi \) for skewed outcomes.

### 3.3 Assessing the positivity assumption

Positivity is the only identification assumptions that can be assessed empirically. The assumption requires that the probability of treatment is bounded \( 0 < P(A = 1 | L) < 1, \forall L \). Violations of positivity (eg, \( P(A = 1 | L) = 1 \)) imply that there are subgroups of the data for which no comparator patients exist, thus forcing the model to extrapolate when computing causal contrasts. Incorrect extrapolation in these regions will bias causal effect estimates. There are many methods of handling violations once they are identified (Petersen et al., 2012), but these are out of scope for this paper. Here we simply provide a framework for assessing this assumption within the unique context of our zero-inflated DP model. Note in Equation (1), we have explicitly modeled treatment probability as a function of covariates. This allows us to predict treatment probability for each patient, given posterior draws \( (\omega^{(t)}_{1:n})_{t=1:T} \) via Monte Carlo:

\[
P(\tilde{A} = 1 | l, D) \approx \frac{1}{T} \sum_{t=1}^T \frac{a}{a+\alpha} \int_\omega p(\tilde{A} | l, \omega)p(l | \omega)dG_0(\omega) + \frac{1}{a+\alpha} \sum_{t=1}^T \int_\omega p(\tilde{A} | l, \omega^{(t)}_i)p(l | \omega^{(t)}_i)
\]

\[
+ \frac{a}{a+\alpha} \sum_{t=1}^T \int_\omega p(l | \omega^{(t)}_i)dG_0(\omega) + \frac{1}{a+\alpha} \sum_{t=1}^T p(l | \omega^{(t)}_i).
\]

A derivation is provided in Web Appendix A in the supplement. Using the above, we can compute \( P(\tilde{A} = 1 | L_i, D) \) for each subject in our sample. Typically, histograms of these probabilities are plotted for treated and untreated patients separately. Separated distributions indicate a lack of overlap and, therefore, high posterior belief of a positivity violation.

### 4 SIMULATION STUDY

In this section, we evaluate bias of \( \Psi_{\text{BNP}} \), coverage of the credible interval estimates, and precision of the estimate as measured by interval width. We compare results to existing methods that may be considered by researchers faced with zero-inflated outcomes—namely BART, a non-Bayesian doubly robust estimator, and two parametric Bayesian gamma models. BART is a BNP, tree-based ensemble for
the conditional mean of the outcome. The doubly robust estimator is a two-part model for treatment assignment and the outcome. We use a boosted frequentist logistic regression for the treatment model and a frequentist Gaussian model for the outcome. The first parametric model is a Bayesian gamma hurdle model. This is a two-part model that explicitly models the probability of the outcome being zero with a logistic regression, while modeling positive outcomes with a gamma regression. The second parametric model is a naive, yet somewhat common, approach of adding .01 to zero outcome values and modeling this transformed outcome using a Bayesian gamma regression. We refer to this as the gamma +.01 model.

We simulate from two data-generating processes (DGPs). In the clustered DGP, we simulate data from three distinct clusters—each with its own set of parameters that govern confounder distributions, binary treatment assignment, zero inflation, and gamma-distributed positive outcomes. The gamma distribution is used to simulate realistic cost data that are nonnegative and skewed within each cluster. Thus, the local conditional outcome distribution assumed in Equation (1) is deliberately misspecified. In the parametric DGP, we simulate data from a single cluster with a common covariate distribution, treatment assignment model, zero inflation, and gamma-distributed positive outcomes. The data are skewed, but not multimodal. Other simulation details regarding hyperparameter settings and sampling are given in Web Appendix D in the supplement.

Simulation results are presented in Table 1. In the clustered setting, the zero-inflated DP model produces effect estimates with the smallest bias, −8.1% of the true value. The 95% credible interval has close to nominal coverage of 94.2%. Though the gamma hurdle model cannot handle multimodality, it outperforms both BART and the doubly robust estimators due to its explicit modeling of structural zeros and skewness. The latter two models capture neither zero inflation nor multimodality and consequently perform poorly.

In the parametric setting, the zero-inflated DP model again exhibits low bias and close to nominal coverage. BART and the doubly robust models have lower bias, but exhibit slight overcoverage (about 96%) in the interval estimates. The gamma hurdle model is correctly specified in the parametric DGP and so performs the best—exhibiting the lowest bias of 1.4%, 95.1% coverage, and yielding the shortest interval. Relative to this correctly specified hurdle model, the zero-inflated DP has only a slightly wider interval length on average (22 034.1 vs 21 778.7), suggesting little efficiency loss. BART and the doubly robust estimators both have wider intervals than the DP on average.

The particularly bad performance of the naive gamma +.01 model—under both DGPs—should be noted. While it is a simple, seemingly harmless trick, adding a small constant severely degrades the accuracy and precision of treatment effect estimates. Unlike the hurdle model and DP model, it does not model structural zeros, and so ignores the effect of treatment that generates these zeros. The zero-inflated DP mixture captures multimodality, skewness, and the treatment’s effect on structural zeros. This allows for good treatment effect estimates under both simple and pathological data distributions with minimal efficiency loss if the parametric model is correct.

### Table 1 Results across 1000 simulated datasets with 3000 subjects each

| DGP          | Model          | Bias  | Coverage | Interval width |
|--------------|----------------|-------|----------|----------------|
| Clustered    | Zero-inflated DP | −0.081 | 94.3%    | 21 612.2       |
|              | BART            | −0.746 | 76.2%    | 26 374.2       |
|              | Doubly robust   | 0.795  | 87.1%    | 33 449.3       |
|              | Gamma hurdle    | −0.509 | 79.8%    | 19 692.2       |
|              | Gamma +.01      | 1.817  | 4.7%     | 27 358.1       |
| Parametric   | Zero-inflated DP | 0.097  | 95.1%    | 22 034.1       |
|              | BART            | −0.054 | 96.1%    | 23 825.3       |
|              | Doubly robust   | −0.027 | 95.9%    | 23 339.1       |
|              | Gamma hurdle    | −0.014 | 95.1%    | 21 778.7       |
|              | Gamma +.01      | −0.489 | 100%     | 50 580.3       |

Average bias of the posterior mean is reported as a proportion of the true value (Ψ = −9740.3 in the clustered setting and Ψ = −10 148.4 in the parametric setting). Mean credible interval widths are presented for the zero-inflated DP model, the BART model, and two gamma models. Confidence intervals are given for the doubly robust method. In the parametric setting, we have 45% in the clustered setting and 55% in the parametric setting. We simulate with one continuous covariate and four binary covariates, all of which affect zero-inflation, treatment probability, and the outcome. All models condition on the simulated confounders so that ignorability holds.

### 5 | APPLICATION: INPATIENT MEDICAL COSTS FOR ENDOMETRIAL CANCER

In this section, we use the proposed DP mixture of zero-inflated regressions to analyze inpatient medical costs among patients with endometrial cancer. Patients who were diagnosed with endometrial cancer between 2000 and 2014 were identified in the SEER Medicare database. Those assigned to either radiation or chemotherapy posthysterectomy were followed for a maximum of 2 years after initial treatment. The total inpatient costs, measured in 2018 US dollars, accrued over the follow-up period was recorded and is our primary outcome of interest. Inpatient costs are costs that accrue during overnight hospitalizations and do not include costs such as prescription treatment costs, outpatient costs, or hospice care costs.

Table 2 presents baseline characteristics of the two treatment groups. There is a significant proportion of zero costs—15.2% in the chemotherapy arm versus 7.9% in the radiation...
TABLE 2 Baseline characteristics: Means and standard deviations are reported for continuous variables

|                     | Chemotherapy (N = 92) | Radiation (N = 952) | SMD  |
|---------------------|-----------------------|---------------------|------|
| Total inpatient costs ($) | 22.1 (28.6)          | 23.4 (34.5)         | 0.039|
| Zero costs          | 14 (15.2%)            | 75 (7.9%)           |      |
| Age (years)         | 73.68 (6.98)          | 73.25 (5.98)        | 0.066|
| Household income ($) | 64.4 (32.4)           | 56.8 (26.2)         | 0.257|
| White               | 76 (82.6%)            | 835 (87.8%)         | 0.147|
| Diabetic            | 20 (21.7%)            | 197 (20.7%)         | 0.026|
| CCI                 | 0.350                 |                     |      |
| 0                   | 49 (53.3%)            | 529 (55.6%)         |      |
| 1                   | 22 (23.9%)            | 260 (27.3%)         |      |
| ≥2                  | 21 (22.8%)            | 131 (13.8%)         |      |
| Grade = 1           | 28 (30.4%)            | 208 (21.8%)         | 0.196|
| FIGO stage I-N0 or I-A | 63 (68.5%)           | 357 (37.5%)         | 0.653|

Counts and percentages are reported for categorical variables. Standardized mean differences (SMD) are provided. All monetary amounts are in thousands of 2018 U.S. Dollars.

arm. Chemotherapy subjects have lower inpatient costs over the follow-up period. However, there may be several confounding factors. For example, the primary determinants of posthysterectomy treatment are the stage and grade of the cancer, with consideration for patient comorbidity and age. These factors, which are measured pretreatment, likely also affect inpatient costs. The standardized mean difference for stage, grade, and Charlson comorbidity index (CCI) are all > 0.1.

In the following subsections, we demonstrate how our method can be used to model the data from several angles. All results are from posterior sampling of the model in Equation (1). We control for race, CCI, household income, cancer grade, and stage in both the positive outcome model and the zero-probability model. We model treatment assignment as a function of these confounders as well. We assume local Gaussian distributions for CCI and household income and Bernoulli distributions for binary covariates. Details about hyperparameter settings, priors, and sampling results are provided in Web Appendix E in the supplement.

5.1 | Multimodality and clustering results

The patients in this study are heterogeneous in terms of their observed costs and covariates. Some have extremely high costs, more comorbidities, and come from varying socioeconomic backgrounds. Clustering can be useful for both describing these groups in terms of observed characteristics or motivating new research. There is a vast literature on clustering methods, and we do not claim the DP-induced method is superior, but it does have several advantages. First, since the DP mixture assumes there are infinitely many clusters in the population (though in a particular analysis, the number of clusters is bounded by $n$), we need not specify the number of clusters beforehand. Second, this method allows for uncertainty quantification around the posterior mode data partition.

Two potentially important confounders of costs and treatment assignment are household income and CCI. The first two panels of Figure 1 visualize cost along these dimensions. While we initialize the model with five clusters, the model identified 10 clusters in the posterior—introducing five additional clusters to accommodate the complexity of the data. In

FIGURE 1 Clustering results from the zero-inflated DP mixture. Colors indicate posterior mode cluster assignment. The first panel projects clustering results onto the cost-income space, and the second projects the results onto the cost-CCI space—both relevant dimensions for understanding costs. The third panel visualizes the full posterior mode matrix discussed in Section 2.3 with a network diagram. Each node represents a patient and the lengths of vertices connecting any two nodes are inversely proportional to the posterior probability of being clustered together. The position of the nodes in x-y space has no meaning (hence the absence of axis labels), only the relative distance between nodes is relevant.
the first row, we see the orange cluster has very high costs, the blue cluster has moderately high costs, while the green and red clusters have lower costs. There are two results worth noting. First, the light blue and gray clusters represent patients who have such distinctly high costs that the DP model places them in their own cluster. Second, the black points represent patients who, while having similar costs to most patients, have distinctly high household income. Thus the DP model places them in their own cluster. From this we can see that the clustering is happening in multiple dimensions rather than only on the cost space.

Similarly, we cannot see much difference between the green and red clusters on the cost-household income space. However, the second panel shows that these patients occupy distinct places on the cost-CCI space, where the red cluster ranks lower than green on CCI. It may be clear at this point that visualizing clustering in two dimensions is limited by the need to choose the variables on each dimension. The third panel solves this issue by visualizing the entire posterior mode matrix discussed in Section 2.3 as a network diagram. We can use this diagram to get a sense of the uncertainty around the mode cluster assignment/partition. For example, the nodes between the red and green clusters have very uncertain assignment. About half the time, they were clustered with the red patients and the other half they were clustered with the green patients—indicating we should not have very high confidence in their posterior mode assignment. This type of uncertainty characterization is absent in many classical clustering algorithms, like K-means. We can summarize observed characteristics of patients by posterior mode assignment. In the orange cluster, average cost in this cluster is $71,139. The distribution of CCI in this group is skewed much higher, suggesting a possible positive association between cost and CCI. On the other hand, we can see from Figure 1 that the light blue cluster has much higher costs (first panel), yet these subjects are relatively low on the CCI scale (second panel). The relationship between cost and CCI seems unclear—and perhaps this motivates future research targeted at learning this relationship.

5.2 Cost prediction in the presence of zero inflation

Induced clustering is the core strength of DP mixtures: A single parametric model estimated using heterogeneous data will have worse fit than an ensemble of locally parametric models fit on more homogeneous partitions. Figure 2 demonstrates the proposed model’s effectiveness at capturing the cost distribution. The predictive cost distributions are quite similar to the observed distribution. This is not the case for the BART and hurdle models—which fail to capture the high end of the distribution and, therefore, consistently underpredict costs. In the second row of plots, we see the DP model occasionally predicts very high costs, while having the bulk of the predictions at < $50,000. Both BART and the hurdle model capture the lower end of the cost distribution well—also predicting the bulk of the costs at < $50,000. However, they rarely predict costs at the high end—thus, failing to capture skewness.

5.3 Estimating causal contrasts and assessing overlap

Finally, we use our method to estimate differences in costs that would have accumulated over 2 years under hypothetical interventions where everyone received radiation versus everyone received chemotherapy as their first posthysterectomy treatment. After applying the method of Section 3.3, Figure 3 shows there is adequate overlap between the two treatment groups, reducing concerns about positivity violations. We use the method of Section 3 to compute a marginal causal effect, a median causal effect, and a risk ratio contrasting the probability of zero cost under radiation versus chemotherapy. Posterior means and credible intervals are displayed in Table 3. Under standard causal identification assumptions, we estimate the causal difference in costs to be $1,672 (Credible Interval: −2,566, 5,722), showing radiation therapy to be more expensive. We estimate a median causal difference to be $872 (CI: −833, 2,790). Finally, we estimate that the probability of having zero costs under radiation therapy is 50% (CI: 0.31, 0.78) lower than under chemotherapy. These results are consistent with unadjusted results (see Table 2).

Marginal causal effects from BART and the gamma hurdle model are roughly in-line with the DP mixture estimates but suffer from relative ineffectiveness at predicting high costs, as explained in the previous section. The risk ratio estimate from the hurdle model is similar to the DP estimate. We note that, consistent with simulation results, the marginal causal effect estimate from the gamma +.01 model differs greatly from the other three models.

6 Discussion and future work

The proposed DP mixture is ideal for capturing joint distributions with continuous, zero-inflated outcomes. It is multipurpose: simultaneously modeling structural zeros, inducing clustering to handle multimodality, and accommodating skewness in the outcome and covariates. As we show in our simulation studies, these traits allow our proposed DP mixture to both produce high-quality causal effects estimates as well as capture the entire outcome distribution. At the same time, posterior draws from the model can be used to perform posterior checks evaluating the validity of positivity.
FIGURE 2  Top row: QQ plots of percentiles (0.02 to 0.98 in increments of 0.02) of the observed cost distribution against predictive cost distributions. Each gray line is a draw of the same size as the data from the predictive cost distribution. The blue line indicates the mean of each percentile across these predictive draws, while the dashed line indicates equality (a perfect fit). The DP mixture opens new cluster to capture skewness—resulting in a predictive distribution closely matching the observed data. The BART model and hurdle model cannot capture this extreme skewness. This is also demonstrated in the bottom row: The DP model occasionally predicts very high costs, while predictions from BART and hurdle models hardly ever predict such high costs.

FIGURE 3  Posterior propensity scores calculated using Equation (5) for both groups indicate adequate overlap—suggesting no evidence of positivity violations.
One might expect BART to perform better than it did in our simulations and in our analysis of cancer data. After all, BART is said to be “effectively nonparametric” (Chipman et al., 2010). However, while it flexibly models the conditional outcome mean, BART still assumes that the outcome distribution is Gaussian—yielding biased estimates in simulation settings where the data are drawn from a skewed distribution like the gamma. Moreover, it does not account for multimodality as it assumes the data are generated from a single mean function and single error variance. The DP mixture makes no such assumptions—yielding better estimation of the entire outcome distribution, as was shown in Figure 2. We note that very recently George et al. (2018) proposed extending BART by modeling the error term nonparametrically using a DP mixture. This may better equip BART to handle skewness, though multimodality will likely remain challenging.

Finally, we consider several extensions of our model for future work. First, while our model provides a framework for assessing positivity, designing a solution within the framework of our model is an important extension. Second, unmeasured confounding is always a concern in observational studies. In our application, this concern is mitigated by the fact that posthysterectomy treatment assignment mechanism is well defined by American Cancer Society guidelines to be mostly driven by patient comorbidity, age, cancer stage, and grade. In many settings, the treatment mechanism may be less well-understood, necessitating sensitivity analyses. Third, standardization provides valid causal estimates only in scenarios with time-constant treatment and confounding. Extending this model to a setting with time-varying confounding would be a worthwhile endeavor.

ACKNOWLEDGMENTS

We used the linked SEER-Medicare database and acknowledge the efforts of the Applied Research Program; National Cancer Institute; Office of Research, Development and Information; Centers for Medicare and Medicaid Services; Information Management Services; and SEER program tumor registries in the creation of the SEER-Medicare database. This work was partly supported by grants R01GM112327 and 124268-IRG-78-002-35-IRG. We thank Dr. Emily M. Ko for data access and clinical guidance. We thank the editors and reviewers for thoughtful revisions that greatly improved our manuscript.

**DATA AVAILABILITY STATEMENT**

The data that support the findings of this study are available from SEER-Medicare. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://healthcaredelivery.cancer.gov/seeermicare/ with the permission of SEER-Medicare.

**ORCID**

Arman Oganisian https://orcid.org/0000-0002-0437-4611
Nandita Mitra https://orcid.org/0000-0002-7714-3910
Jason A. Roy https://orcid.org/0000-0003-4237-0504

**REFERENCES**

Barcella, W., De Iorio, M., Baio, G. and Malone-Lee, J. (2016) A Bayesian nonparametric model for white blood cells in patients with lower urinary tract symptoms. *Electronic Journal of Statistics*, 10, 3287–3309.

Blackwell, D. and MacQueen, J.B. (1973) Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1, 353–355.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266–298.

Dahl, D.B. (2006) Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model, Cambridge, UK: Cambridge University Press, pp. 201–218.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209–230.

George, E., Laud, P., Logan, B., McCulloch, R. and Sparapani, R. (2019) Fully nonparametric Bayesian additive regression trees. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part B (Advances in Econometrics)*, 40, 89-110.

Ghosh, S.K., Mukhopadhyay, P. and Lu, J.-C. (2006) Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136, 1360–1375.

Hannah, L.A., Blei, D.M. and Powell, W.B. (2011) Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12, 1923–1953.

Hill, J.L. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20, 217–240.

Keil, A.P., Daza, E.J., Engel, S.M., Buckley, J.P. and Edwards, J.K. (2017) A Bayesian approach to the g-formula. *Statistical Methods in Medical Research*, 27, 3183–3204.

Kim, C., Daniels, M.J., Marcus, B.H. and Roy, J.A. (2017) A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics*, 73, 401–409.
Kurz, C.F., Hatfield, L.A. (2019). Identifying and interpreting subgroups in health care utilization data with count mixture regression models. *Statistics in Medicine*, 38, 4423–4435.

Linero, A.R., Sinha, D. and Lipsitz, S.R. (2020) Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76, 131-144.

Muller, P., Quintana, F., Jara, A. and Hanson, T. (2015) *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Cham, Switzerland: Springer International Publishing.

Neal, R.M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249–265.

Petersen, M.L., Porter, K.E., Gruber, S., Wang, Y. and van der Laan, M.J. (2012) Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21, 31–54.

Rodriguez, C.E. and Walker, S.G. (2014) Label switching in Bayesian mixture models: deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23, 25–45.

Roy, J., Lum, K.J. and Daniels, M.J. (2017) A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18, 32–47.

Roy, J., Lum, K.J., Zeldow, B., Dworkin, J.D., Re, V.L., III, and Daniels, M.J. (2018) Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, 74, 1193–1202.

Rubin, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6, 34–58.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.

Wade, S., Dunson, D.B., Petrone, S. and Trippa, L. (2014) Improving prediction from Dirichlet process mixtures via enrichment. *Journal of Machine Learning Research*, 15, 1041–1071.

Xu, D., Daniels, M.J. and Winterstein, A.G. (2018) A Bayesian nonparametric approach to causal inference on quantiles. *Biometrics*, 74, 986–996.

Xu, Y., Müller, P., Wahed, A.S. and Thall, P.F. (2016) Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association*, 111, 921–950.

**SUPPORTING INFORMATION**

Web Appendices, Tables, and Figures referenced in Sections 1-5 are available with this paper at the Biometrics website on Wiley Online Library. Relevant implementation code is also available.

**How to cite this article:** Oganisian A, Mitra N, Roy JA. A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics*. 2021;77:125–135. [https://doi.org/10.1111/biom.13244](https://doi.org/10.1111/biom.13244)