Bayesian Nonparametric Common Atoms Regression for Generating Synthetic Controls in Clinical Trials

Noirrit Kiran Chandra, Abhra Sarkar, John F. de Groot, Ying Yuan, and Peter Müller

Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX; Department of Statistics and Data Sciences, The University of Texas at Austin, TX; Department of Neurological Surgery, University of California San Francisco, San Francisco, CA; Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX; Department of Mathematics, The University of Texas at Austin, TX

ABSTRACT
The availability of electronic health records (EHR) has opened opportunities to supplement increasingly expensive and difficult to carry out randomized controlled trials (RCT) with evidence from readily available real world data. In this article, we use EHR data to construct synthetic control arms for treatment-only single arm trials. We propose a novel nonparametric Bayesian common atoms mixture model that allows us to find equivalent population strata in the EHR and the treatment arm and then resample the EHR data to create equivalent patient populations under both the single arm trial and the resampled EHR. Resampling is implemented via a density-free importance sampling scheme. Using the synthetic control arm, inference for the treatment effect can then be carried out using any method available for RCTs. Alternatively the proposed nonparametric Bayesian model allows straightforward model-based inference. In simulation experiments, the proposed method exhibits higher power than alternative methods in detecting treatment effects, specifically for nonlinear response functions. We apply the method to supplement single arm treatment-only glioblastoma studies with a synthetic control arm based on historical trials. Supplementary materials for this article are available online.

1. Introduction
We introduce a novel Bayesian nonparametric regression model to construct synthetic control arms from external real world data (RWD) to supplement single arm treatment-only trials. The use of common atoms across multiple random probability measures is a critical feature of the proposed construction. Models with similar features have been used before in the literature, including Denti et al. (2021), Camerlenghi et al. (2019), Rodriguez, Dunson, and Gelfand (2008), and Teh et al. (2006).

Randomized controlled trials (RCT) are the gold standard in evidence-based evaluation of new treatments. RCTs are, however, increasingly associated with bottlenecks involving volunteer recruitment, patient truancy and adverse events (Nichol, Bailey, and Cooper 2010) and hence are often very time consuming, expensive and laborious. This is of particular concern for rare diseases, such as glioblastoma (GBM). With digitization of health records and other advances in medical informatics, new data sources are becoming available that can supplement RCTs. For example, relevant information on a control treatment is often available from completed RCTs, electronic health record data, insurance claims data or patient registries from hospitals (Franklin et al. 2019). Such external data, also referred to as RWD, can augment or substitute the control group in the target clinical trial (Davi et al. 2020). This has led researchers to consider the creation of synthetic control arms from RWD (see Schmidli et al. 2020 for a review). However, the heterogeneity of RWD prohibits the direct use of patient level data as a control arm, lest differences with the actual treatment population with respect to patient profiles bias inference on treatment effects (Burcu et al. 2020). Many existing methods adjust for the lack of randomization in treatment assignments by correcting the bias in the response model and hence can be sensitive to the specification of the treatment assignment as well as the response model as we discuss below. In this article, we take a fundamentally different approach by resampling the RWD to construct a cohort equivalent to the treatment arm in terms of their covariate profiles which can then serve as the (synthetic) control arm.

There is a fast growing literature on the problem of incorporating RWD in clinical trials. Traditional meta-analytic approaches aim to combine information across studies to construct comparisons of treatments (Sutton and Abrams 2001). Power prior (Prevost, Abrams, and Jones 2000; Chen and Ibrahim 2000), commensurate prior (Hobbs et al. 2011) and elastic prior (Jiang, Nie, and Yuan 2023) constructions try to incorporate information from historical data by way of informative prior models. However, these approaches may be inadequate when the RWD population is considerably more heterogeneous than the experimental arm; see Müller, Chandra, and Sarkar (2023) for a review.

Many methods to incorporate RWD in trial design and data analysis are based on propensity scores (PSs), defined as the conditional probabilities of treatment assignment given...
covariates. In the context of incorporating external data, investigators often use PSs for a patient being selected into the current trial versus the external data; in case of supplementing a single arm treatment-only trial, the PSs are identical to treatment assignments. Rosenbaum and Rubin (1983) showed that an unbiased estimate of the average treatment effect can be obtained by PS adjustments. Most PS-based methods can be broadly classified to be based on matching, stratification, weighting, or regression. Matching is used to achieve covariate balance across different arms. However, matching PSs do not generally imply matching covariates (King and Nielsen 2019). Stratification splits the data into strata with respect to PSs and calculates an average treatment effect as a weighted average of within-stratum estimates (Wang et al. 2019; Chen et al. 2020; Lu et al. 2022). PS-stratification may be sensitive to the definition of the strata and weight-based estimators may be sensitive to the misspecification of the PS model (Zhao 2004). Regression adjustments, that use the PS as a regressor for the outcome, address these issues (Rosenbaum and Rubin 1983) but the estimates may again be biased if the regression model is misspecified (Vanderweele and Robins 2014). Bayesian nonparametric models that avoid a particular parametric family or structure, such as linearity, of the regression relationship have thus also been proposed (Wang and Rosner 2019). Nevertheless, consolidated unidimensional PSs can be inadequate in matching multivariate covariates from multiple studies (Stuart 2010; King and Nielsen 2019). Additionally, these methods often do not efficiently use all available data by dropping unmatched data. Finally, some other methods (Hasegawa et al. 2017; Li and Song 2020), although not specifically designed to create synthetic controls, also integrate multiple studies using the covariate distributions.

In this article, we develop an alternative approach based on Bayesian nonparametric (BNP) mixture models. Mixture models imply a random partition of experimental units linked to different atoms in the mixture (Dahl 2006). We exploit this property to propose a BNP common atoms mixture (CAM) model to introduce matched clusters of patients in a treatment-only trial dataset and a (typically much larger) RWD. We show how such matched clusters allow a density free importance resampling scheme to generate a subpopulation of the RWD such that the distribution of covariates in the subpopulation can be considered to be equivalent to the single-arm trial. That is, the patients in a matching RWD cluster can be considered digital clones of patients in a matching cluster in the single-arm trial.

The proposed CAM model allows, among other things, the following two alternatives for inference on treatment effects. Having established equivalent patient populations, inference can in principle proceed as if treatment had been assigned at random, using inference for RCTs. Alternatively, we propose model-based inference using an extension of the CAM model with a sampling model for the outcome. While both alternatives are based on the same underlying CAM model, we prefer the model-based inference on treatment effect as a more explicit and principled approach.

The proposed CAM model builds on related BNP models in the literature, including the hierarchical Dirichlet process (DP) (Teh et al. 2006) which allows for information sharing across multiple groups through common atoms, the nested DP (Rodriguez, Dunson, and Gelfand 2008) which can identify distributional clusters, and Camerlenghi et al. (2019) who proposed a latent mixture of shared and idiosyncratic processes across the sub-models. Denti et al. (2021) proposed a CAM model for the analysis of nested datasets where the distributions of the units differ only over a small fraction of the observations sampled from each unit. In contrast to these constructions, the CAM model proposed here introduces more structure as needed in our application by setting up two nonparametric Bayesian mixture models with shared atoms and constraints on the implied clusters.

The rest of this article is organized as follows. Section 2 describes the glioblastoma study that motivated this work. Section 3.1 introduces the proposed common atoms mixture model on the covariates and how it can handle variable dimensional covariates of different data types; Section 3.2 introduces a novel density-free importance resampling scheme to achieve equivalent populations; and Section 3.3 discusses the general common atoms regression model, a flexible mixture of lognormals for censored survival outcomes and an easy to use graphical tool for model validation. In Section 4, we discuss two alternative strategies for inference on treatment effects. Section 6 presents simulation studies. Section 7 shows results for the motivating GBM data. Section 8 concludes with final remarks. Below, in Table 1, we list the many acronyms used in the article for easy reference.

### 2. Motivating Application in Glioblastoma

Our motivating application arises from a GBM data science project at MD Anderson Cancer Center. GBM is a devastating disease with the average life expectancy of less than 12 months in the general population (Ostrom et al. 2016). Despite decades of intensive clinical research, the progress in developing an effective treatment for GBM lags behind that of other cancers (Aldape et al. 2019). In the last 30 years, only two drugs (carmustine wafers and temozolomide) have been approved by the Federal Drug Administration (FDA) for patients with newly diagnosed GBM (Fisher and Adamson 2021). These drugs extend median survival by less than three months and neither offers a potential for cure. One major cause of the high failure rate of the drug development for GBM is suboptimal design of phase II trials, in particular, the lack of a control arm in many studies (Grossman et al. 2017). A review of phase I/II GBM trials from 1980 to

![Table 1. List of acronyms.](image)

| Acronym | Full forms | Acronym | Full forms |
|---------|-----------|---------|-----------|
| AUC     | Area under the receiver operating characteristic curve | DP | Dirichlet process |
| BART    | Bayesian additive regression tree | GBM | Glioblastoma |
| BNP     | Bayesian nonparametric mixture | IS | Importance sampling |
| CAM     | Common atoms mixture | PPMx | Product partition model with regression on covariates |
| CA-PPMx | Common atoms PPMx | PS | Propensity score |
| RCT     | Randomized controlled trial | RWD | Real-world data |
2013 found that only 20 (5%) were randomized compared to 365 (95%) single-arm trials (Grossman and Ellsworth 2016). Reasons for the dominance of single-arm trials include the small number of GBM patients available for clinical trials and investigator’s desire to speed up drug development and reduce trial costs. GBM is a rare disease by the definition of the Orphan Drug Act (FDA 2020). Unfortunately, the high heterogeneity of GBM patients makes single-arm trials highly susceptible to bias, contributing to the fact that almost all phase II trials showing promising treatment effects failed in phase III RCTs (Mandel et al. 2017). The objective of the GBM data science project is to address this pressing issue by leveraging historical data collected at the MD Anderson Cancer Center. The overarching goal is to develop a platform for future single-arm clinical trials in GBM, with synthetic controls constructed from the historical database to enhance the evaluation and screening of new drugs. Working toward this goal, we describe here a method to create synthetic controls, as the engine of the platform, for future trials.

We work with a database that comprises records from 339 highly clinically and molecularly annotated GBM patients treated at MD Anderson over more than 10 years. Once the system is set, the database is expected to be continuously updated with new patient data collected at MD Anderson Cancer Center and potentially also be combined with the data from other institutions.

After discarding variables with minimal variability across patients and relying on clinical judgment, we identified 11 clinically important categorical covariates. These covariates are commonly considered as prognostic factors in GBM treatments (Nam and de Groot 2017; Alexander et al. 2019) and are briefly described in Table 2.

Figure 1 shows the categorical covariates in the historical database and a future treatment-only study which we elaborate in Section 7. Figure S.1 in the supplementary materials highlights the lack of randomization in the two populations.

3. Common Atoms Mixture Model

We first introduce a model for matching patients with respect to their covariate profiles across different treatment arms and then an extension of the model to also include outcomes. Later we will introduce two alternative methods for inference on treatment effects that build on this model.

### 3.1. Common Atoms Mixture Model on the Covariates

Suppose we have $S$ datasets $(X_{si}, Y_{si}), s = 1, 2, \ldots, S$, comprising $p$-dimensional covariate vectors $X_{si} = (X_{si,1}, \ldots, X_{si,p})^T$ and corresponding responses $Y_{si}$ associated with patients $i = 1, \ldots, n_s$. In this article, we assume the responses to be univariate. Let $s = 1$ refer to the arm for the (new) experimental therapy, and $s = 2, \ldots, S$ denote the RWD datasets. Focusing on the motivating GBM application, we elaborate the model for $S = 2$ with a single RWD set. When we have multiple historical datasets, that is, when $S > 2$, we would simply merge them and consider the merged dataset to be a single RWD with increased heterogeneity as illustrated in Section S.9.6 of the supplementary materials. For a valid evaluation of treatment effects, it is then important to verify equivalent patient populations, that is, matching the distributions of $X_{si}$ under $s = 1$ versus $s = 2$, or to otherwise adjust for any detected differences (Burcu et al. 2020). As the RWD population can be from a variety of sources,

| Covariate | Description |
|-----------|-------------|
| Age       | Dichotomized at 55 years |
| KPS       | Karnofsky performance score, categorized into three classes: “≤ 60”, “(60, 80]” and “> 80” |
| RT Dose   | Radiation therapy dose: dichotomized at 50 Gray |
| SOC       | Received standard-of-care (concurent radiation therapy and temozolomide): Yes/No |
| CT        | Participation in a therapeutic trial: Yes/No |
| MGMT      | Status of MGMT (O6-methylguanine-DNA methyltransferase) gene: methylated (M), unmethylated (UM) or uninterpretable (UI) |
| ATRX      | Loss of the ATRX chromatin remodeler gene: Yes/No |
| Gender    | Gender |
| EOR       | Extent of tumor resection: “total”, “subtotal” or “laser interstitial thermal therapy” (LITT, Patel and Kim 2020) |
| Histologic grade | Grade of astrocytoma/IV (GBM) (most cases), or I–III (low-grade or anaplastic) (few) |
| Surgery reason | “therapeutic” or “other” (relapse) |

Figure 1. Glioblastoma dataset of 11 baseline categorical covariates with missing entries in the two treatment arms. The left block shows the historical patients. The (smaller) right block shows a hypothetical future trial.
such data are typically more heterogeneous than the patient population in the ongoing trial. We develop a novel BNP CAM model with this specific feature to model the two distributions. The proposed CAM model gives rise to a random partition of similar $X_{2,i}$ and a matching partition of $X_{2,j}$. Clusters under the latter partition can be considered digital clones of the matching clusters of the earlier partition.

We first construct the model for covariates $X_{2,i}$ in the RWD. Let $\tilde{\xi} = \{\tilde{\xi}_j\}_{j=1}^{\infty}$ and $\pi_2 = \{\pi_2j\}_{j=1}^{\infty}$ denote cluster-specific parameters and weights, respectively. We let

$$X_{2,i} \mid \tilde{\xi}, \pi_2 \sim \sum_{j=1}^{\infty} \pi_2j q(X_{2,i} \mid \tilde{\xi}_j),$$

$$\tilde{\xi}_j \mid \xi \sim G_0(\tilde{\xi}_j \mid \xi),$$

$$\pi_2 \sim \text{GEM}(\alpha_2).$$

(1)

Here $q(\cdot \mid \tilde{\xi}_j)$ is a suitably chosen kernel with parameter $\tilde{\xi}_j$, $G_0(\cdot \mid \xi)$ is a prior distribution for the $\tilde{\xi}_j$'s, and GEM$(\alpha)$ is a stick-breaking prior on the mixture weights corresponding to a DP with mass parameter $\alpha > 0$ (Sethuraman 1994). Let $G = \sum_{j=1}^{\infty} \pi_2j \delta_{\tilde{\xi}_j}(\cdot)$ denote a discrete probability measure with atoms at the $\tilde{\xi}_j$'s. An equivalent hierarchical model representation of (1) is

$$X_{2,i} \mid \xi \sim q(X_{2,i} \mid \xi),$$

$$\xi \mid G \sim G,$$

$$G \mid \alpha_2, \xi \sim \text{DP}(\alpha_2, G_0(\cdot \mid \xi)),$$

(2)

where DP$(\alpha, G_0)$ is a DP with base measure $G_0$ and concentration parameter $\alpha$ (Ferguson 1973). The discrete nature of the DP random measure $G$ gives rise to possible ties between the $\xi_j$'s, which define the desired clusters. For later reference we define notations for these ties and clusters. Let $\xi^* = \{\xi^*_j, j = 1, \ldots, k(n_2)\}$ denote the distinct values in $\{\xi_j, i = 1, \ldots, n_2\}$, let $c_2,j = j$ if $\xi_i = \xi^*_j$ denote cluster membership indicators defining clusters $C_j = \{i : \xi_i = \xi^*_j\}$. We assume the distribution of $X_{2,i}$ be a mixture with the same kernel $q$ and the same atoms $\xi^*$.

$$F_i(X_{1,i} \mid \pi_1(\xi^*)) = \sum_{j=1}^{k(n_2)} \pi_{1,j} q(X_{1,i} \mid \xi^*_j),$$

$$\pi_1 \sim \text{Dir} \left[ \frac{\alpha_1}{\lambda(n_2,1)}, \ldots, \frac{\alpha_1}{\lambda(n_2,n_2)} \right],$$

(3)

where $\pi_1 = (\pi_{1,1}, \ldots, \pi_{1,k(n_2)})$, $\text{Dir}(a_1, \ldots, a_r)$ denotes an $r$-dimensional Dirichlet distribution with parameters $a_1, \ldots, a_r$, and $a_1 > 0$ is a concentration parameter. Note that model (3) is defined conditionally on (1) and $\xi^*$ such that $F_1$ and $F_2$ share the same set of atoms. Importantly, the construction avoids the imputation of clusters (strata) with only $X_{2,i}$'s. There is always a corresponding (non-empty) cluster for the $X_{2,i}$'s from the RWD. This is important for the upcoming constructions. The motivation here is that, owing to the bigger size of the RWD compared to the trial arm, $X_2$ can be expected to exhibit greater heterogeneity than $X_1$ (see, e.g., the right panel in Figure 2).

In summary, we define $F_1(X \mid \pi_1, \xi^*) = \sum_{j=1}^{k(n_2)} \pi_{1,j} q(X \mid \xi^*_j)$ and $F_2(X \mid \pi_2, \tilde{\xi}) = \sum_{j=1}^{\infty} \pi_2j q(X \mid \tilde{\xi}_j)$, with the prior on atoms and weights as discussed. Figure 2 shows a stylized representation of the generative process of the proposed CAM model. Notice that here atom $\tilde{\xi}_3$ is not linked with any $X_2$ observation and hence $k(n_2) = 3$. Accordingly, $F_1(\cdot \mid \pi_1, \xi^*)$ is a mixture of three components. Finally, no observation from $X_1$ is linked to $\tilde{\xi}_2$. The $X_{2,i}$'s linked to $\tilde{\xi}_2$ and $\xi_4$ can be regarded as digital clones of the $X_{1,i}$'s linked to the same atoms.

The described CAM model is different from existing BNP mixture models. In (1)–(3), the atoms linked to $X_1$ are always a subset of those atoms that are linked to $X_2$, which is not naturally the case for the hierarchical DP model (Teh et al. 2006). Also, unlike the nested DP (Rodriguez, Dunson, and Gelfand 2008) and the common atoms nested DP (Denti et al. 2021) models, there is no notion of clustering distributions. That is, $p(F_1(X \mid \pi_1, \xi^*) = F_2(X \mid \pi_2, \tilde{\xi})) = 0$ a priori. Instead, the intention here is to cluster similar covariate values across the datasets.

Regarding the concentration parameters $\alpha_i$, we assume $\log \alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2)$ for $i = 1, 2$. Ascolani et al. (2022) showed that a hyper-prior on the concentration parameters can solve the problem of inconsistency of DP mixtures (Miller and Harrison 2013).

### 3.1.1. Handling Mixed Data Types and Missing Values

An appealing feature of the proposed CAM model over existing approaches is the easy use of covariates of different data-types and missing values. Covariates in RCTs often comprise different data-types including continuous, discrete and categorical variables. Missing values are also quite common. For example, in Figure 1, there are a large number of missing values for the ATRX gene which has only recently been identified as a therapeutic target for glioma (Haase et al. 2018) and was therefore not commonly recorded before.

Many existing methods for handling missing data rely on imputation (Choi, Dekkers, and le Cessie 2019), possibly at the expense of an additional layer of prediction errors. Alternatively, data records with missing variables may be dropped altogether, resulting in a reduced sample size.

Assuming missingness completely at random, the proposed CAM model avoids these issues by accommodating variable dimensional covariates in a principled manner by considering a separate univariate kernel for each covariate. Note that a mixture with independent kernels can still accommodate marginal dependence between the covariates (Ghosal and van der Vaart 2017, sec. 7.2.2, pp. 175). Specifically, let $O_{i,j} = \{i : X_{si,j} is recorded\}$ denotes the set of observed covariates for patient $i$ in dataset $s$. We use independent kernels

$$q(X_{si,j} \mid \xi^* \mid \xi^*_{j,\ell}) = \prod_{\ell \in O_{i,j}} q(X_{si,\ell} \mid \xi^*_{j,\ell}),$$

$$G_0(\xi^*_{j,\ell} \mid \xi) = \prod_{\ell = 1}^{p} g_0(\xi^*_{j,\ell} \mid \xi),$$

(4)

where $q(\cdot \mid \xi^*_{j,\ell})$ is a univariate kernel corresponding to the $\ell$th covariate with parameters $\xi^*_{j,\ell}$ and $g_0(\xi^*_{j,\ell} \mid \xi)$ is a prior on $\xi^*_{j,\ell}$ with hyper-parameters $\xi$. The likelihood function of $X_{si,j}$ is then computed on the basis of only the observed values. The kernel $q_\theta$ is chosen to accommodate the data-type of the $\ell$th covariate. The model allows co-clustering of $X_{si,j}$ with some missing variables and another fully observed $X_{si,j}$; see Section S.3 of the supplementary materials for additional details. Missingness
patterns other than completely at random can be handled by introducing additional hierarchy in the model, see, for example, Linero and Daniels (2018) for a review.

### 3.2. Density-Free Importance Resampling of RWD

Building on the fitted CAM for covariates, we propose an importance resampling method to create a subpopulation of $X_2$ that can be considered to be equivalent to $X_1$ (see below for a definition of equivalence that is being used here). Under the assumption of no unmeasured confounders, the $X_2$'s in the sampled (or weighted) subpopulation can be assumed to follow the same distribution as $X_{1,i}$, and be considered digital clones of the $X_{1,i}$. With such equivalent populations, in principle, any desired method for randomized clinical trials can subsequently be used to carry out inference on treatment effects. Such focus on equivalent populations follows recent recommendations by the FDA (FDA 2021).

Recall that $F_i$ denotes the mixture model for $X_{s,i}$, $s = 1, 2$, under (1) and (3), respectively. We define equivalent populations as a subset (possibly all) of $X_2$ together with a set of weights such that expectation of any function of interest $g(X_{1,i})$ under $X_{1,i} \sim F_1$ can be evaluated as a (weighted) Monte Carlo average using these $X_2$ (and the weights). Here we assume that all stated expectations exist and that the order of taking expectations and limits can be switched.

Recall that $F_2(\cdot | \pi \pi_2, \tilde{\pi}_j) = \sum_{j=1}^{\infty} \pi_2 j q(\cdot | \tilde{\pi}_j)$. Alternatively, the joint model of $(X_2, c_2)$ can be expressed as $F^H_2(\pi \pi_2, c_2 | \pi \pi_2, \tilde{\pi}_j) = q(X_2 | \pi \tilde{\pi}_j) \pi_2 c_2$. For easier housekeeping, we assume $q_j^* = \tilde{\pi}_j$ for $j = 1, \ldots, k(n_2)$, that is, the first $k(n_2)$ atoms are linked with the $X_2$'s. Accordingly, we let $F_1(\cdot | \pi \pi_2, \tilde{\pi}_j) = \sum_{j=1}^{k(n_2)} \pi_2 \pi_j q(\cdot | \tilde{\pi}_j)$ using the same first $k(n_2)$ atoms observed in the $X_2$ population. This is the exact construction of (1) and (3). For an equivalent population, we require weights $w_i$ attached to $(X_2, c_2)$ (using $w_i = 0$ to drop samples) such that:

$$
\mathbb{E}_{F_i(\pi \pi_2, \tilde{\pi}_j)} [g(X_{1,i})] = \mathbb{E}_{F^H_2(\pi \pi_2, c_2 | \pi \pi_2, \tilde{\pi}_j)} [g(X_2, c_2)] \quad \text{with} \\
\tilde{g}(X_2, c_2) = \sum_{i=1}^{n_2} w_i g(X_{2,i}, c_{2,i}).
$$

The weights $w_i$ are functions of $c_{2,i}$ and $\pi_{1,i}$ as follows. Define $n_{2,i} = |C_{2,i}|$, the cardinality of the earlier introduced clusters $C_{2,i}$. Then $\frac{1}{n_{2,i}} \sum_{i \in C_{2,i}} g(X_{2,i})$ is an unbiased estimator of $\mathbb{E}_{g(\pi \tilde{\pi}_j)} g(X)$

$$
\hat{g} = \sum_{j=1}^{\infty} \pi_{1,j} \left\{ \sum_{i \in C_{2,j}} \frac{1}{n_{2,i}} g(X_{2,i}) \right\} = \sum_{i=1}^{n_{2,i}} \frac{n_{2,i}}{n_{2,i}} g(X_{2,i})
$$

is an unbiased estimator of $\mathbb{E}_{F_i(\pi \pi_2, \tilde{\pi}_j)} g(X)$. We then recognize $\pi_{1,i} / n_{2,i}$ as the ideal weights. Since we only observe $X_i$ but not $\pi_1$ and $c_2$, we replace $\pi_{1,i} / n_{2,i}$ in $\hat{g}$ by a Monte Carlo average after posterior MCMC simulation to get the desired equality simulation-exact (i.e., in the limit as $n_1, n_2$ and the number of MCMC simulations increases). Let $m = 1, \ldots, M$ index the posterior sample and use $\pi_{1,m} / n_{2,j}$, etc. to indicate parameter values in the $m$th sample. We use

$$
\tilde{g} = \sum_{j=1}^{\infty} w_j g(X_{2,j}), \quad w_i \propto \sum_{m=1}^{M} \pi_{1,m} n_{2,j} \pi_{1,m} n_{2,j}
$$

with $w_i$ being the importance sampling weight for $X_{2,i}$. The $X_2$'s can be resampled with these weights to obtain the desired subpopulation with distribution $F_j(x | \pi_{1,j} \tilde{\pi}_j)$ (Skare, Belviken, and Holden 2003). This resampled subpopulation of $X_2$ can then be regarded as equivalent in distribution to $X_1$. Algorithm 1 summarizes the procedure.

To test the equivalence of the two populations, we use a Bayesian additive regression tree (BART, Chipman, George, and McCulloch 2010) in Step 5 of Algorithm 1. In extensive simulation studies in Section 6, we notice that an AUC (area under the receiver operating characteristic curve) less than 0.6 yields excellent empirical performance. Once equivalence is achieved, in principle any existing approach for inference on treatment effects can be used (see Section 4 and later).

Note that even if the RWD population is not a heterogeneous superset of the current trial, one can still fit the CAM model.

In the case the RWD is not comparable, Step 5 of Algorithm 1 can discriminate the two populations and the AUC can quantify the degree of incongruence. In general, importance sampling schemes need the ratio of the target density (in our case, $F_j$) and the importance sampling density (in our case, $F_2$). For our problem, this would require high-dimensional density estimation. Even if the densities were known, importance sampling would be plagued by unbounded weights (Au and Beck 2003). Exploiting...
the common atoms structure, our proposed scheme however avoids evaluation of the marginal multivariate densities. We therefore refer to this as a density-free importance resampling scheme, and for brevity often simply as an IS scheme. In the denominator of $w_i$, the use of $n_{2,j}$ (which by definition are $\geq 1$) avoids complications arising from unbounded weights. Conventional importance sampling schemes are asymptotically consistent. This is seen to hold in numerical experiments with our algorithm as well. Additional discussions on Algorithm 1 are in Section S.4 of the supplementary materials.

### 3.3. Regression with CAM Model on Covariates

Note that up to here we only concerned ourselves with the covariates, without any reference to the outcomes $Y$. In preparation for one of the strategies in the upcoming discussion of treatment comparison (Section 4), we now augment the CAM model to include a sampling model for the outcomes. That is, we add a response model on top of the CAM model on covariates.

The extended model defines a regression of $Y_{s,j}$ on covariates $X_{s,j}$ by first grouping patients with similar covariate profiles into clusters and then adding a cluster-specific sampling model for the common atoms structure, our proposed scheme however

The response model (7) depends on the covariates indirectly via $c_{s,j}$'s, that is, the partition induced by the covariates. Within stratum $C_j = C_{1,j} \cup C_{2,j}$, the response model allows for a treatment comparison based on $(\theta_{1,j}, \theta_{2,j})$, which can then be averaged with respect to the assumed distribution of $X_{s,j}$ to define an average treatment effect.

For the implementation in the motivating case study, we let $Y_{s,j}$ denote the log OS (overall survival) times and assume $h(Y_{s,j} | \theta_{s,j})$ to be a normal kernel with $\theta_{s,j} = (\mu_{s,j}, \sigma_{s,j}^2)$. Such mixtures are highly flexible (Ghosal, Ghosh, and Ramamoorthi 1999), making them an attractive choice for many applications. We complete the model with conjugate normal-inverse-gamma (NIG) priors on the $(\mu_{s,j}, \sigma_{s,j}^2)$'s. In summary, we have

$$Y_{s,j} \mid c_{s,j} = j, \theta_{s,j} \overset{\text{iid}}{\sim} N(\mu_{s,j}, \sigma_{s,j}^2), \quad \mu_{s,j} \mid \sigma_{s,j}^2 \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

$$\sigma_{s,j}^2 \overset{\text{iid}}{\sim} \text{Ga}(a_0, b_0),$$

where Ga($a_0, b_0$) is a gamma distribution with mean $a_0/b_0$. We add the hyper-priors $\mu_0 \sim N(m_{\mu}, \sigma^2)$ and $\log b_0 \sim N(m_{\sigma}, \sigma^2)$ on the main location-scale controlling hyper-parameters $\mu_0$ and $b_0$ while fixing the precision hyper-parameters $\kappa_0$ and $a_0$. Choices of these hyperparameters are discussed in Section S.7 of the supplementary materials. Finally, for a goodness-of-fit test under the proposed model, we use the approach of Johnson (2007) to build a graphical tool based on quantile plots. Such visual tools are often quite effective for detecting departures from model assumptions (Meloun and Militý 2011, chap. 2). See Section S.5 in the supplement for more details.

### 4. Inference on Treatment Effects

#### 4.1. Two-Step Importance Sampling (IS) Approach

We already described the use of the weights $\pi_{s,j}$ in the CAM model to achieve equivalent patient populations. This allows a straightforward approach to treatment comparison. Using the adjusted (resampled) subpopulation of $X_{2,1}$, one can proceed with inference on the treatment effect using any method relying on equivalent patient populations across the two arms. We refer to this approach as the “two-step IS” and use it in the simulation studies and applications in Sections 6 and 7, respectively. This approach does not make use of the outcome model of Section 3.3.

#### 4.2. Model-Based Inference for Treatment Effects

Alternatively, we implement inference using the response model of Section 3.3, that is, the full CA-PPMx. We refer to this approach as “model-based inference”. We assume that the desired inference on treatment effects takes the form of inference for some notion of difference $\delta(\cdot, \cdot)$ of the marginal distributions under the two treatment arms, $\Delta_0 = \delta \{ f_1(\cdot) \mid \theta_1, \pi_1 \}, f_2(\cdot) \mid \theta_2, \pi_2 \}$. However, since the covariate populations in the two treatment arms can be substantially different, comparison between the marginal (with respect to the covariates) outcome models $f_1(Y_{1,i} \mid \theta_1, \pi_1)$ and $f_2(Y_{2,i} \mid \theta_2, \pi_2)$ can be biased. We need to appropriately adjust for the
differences in the two populations. We do this by replacing $f_2$ as follows. Exploiting the common atoms structure of the proposed CA-PPMx, there is an operationally simple method to carry out this adjustment and infer treatment effects. Since within each cluster, the covariate populations can be considered equivalent, the adjustment for the lack of randomization amounts to adjusting the corresponding cluster weights. We define

$$\tilde{f}_2(Y \mid \theta_2, \pi_1) = \sum_{j=1}^{k_{\theta_2}} \pi_{1,j} h(Y \mid \theta_{2,j}),$$

where the mixture components $h(Y \mid \theta_{2,j})$ of the response model in the RWD are weighted by $\pi_1$, that is, the cluster weights associated with $X_1$ (rather than $\pi_2$). Thus, $\tilde{f}_2$ is the distribution of outcomes under control in the treatment population or in other words, the response of an average individual from the trial arm potentially treated with the control therapy. With these notions, we define the population adjusted treatment effect as

$$\Delta_0 = \delta[f_1(\cdot \mid \theta_1, \pi_1), \tilde{f}_2(\cdot \mid \theta_2, \pi_1)].$$

For example, when $Y$ is a univariate response variable and $\delta(f_1, f_2) = \mathbb{E}(f_1(Y) - f_2(Y), \tilde{\Delta}_0$ simplifies to

$$\tilde{\Delta}_0 = \sum_{j=1}^{k_{\theta_2}} \pi_{1,j} (\mathbb{E}h(Y_{\theta_{2,j}}(Y) - \mathbb{E}h(Y_{\theta_{2,j}}(Y))).$$

In general, each cluster of covariates in the CAM model can be interpreted as a homogeneous sub-population of patients. For the $j$th group, the average treatment effect is $\delta(h(Y \mid \theta_{2,j}), h(Y \mid \theta_{2,j})$ and its proportion in the target population is $\pi_{1,j}$. The reported treatment effect (9) includes the adjustment with the sub-population proportions $\pi_{1,j}$. On a related point, the proposed model-based inference on treatment effects in the CA-PPMx model can be interpreted as a stochastic propensity score stratification approach. See Section S.6 in the supplementary materials for the details.

We prefer the Bayesian model-based approach to avoid discarding unmatched patient records from the RWD from the analysis. The two-step IS can be useful to validate the results obtained by the model-based approach.

5. Posterior Computation

We develop an efficient Gibbs sampler for posterior inference in the proposed CAM model for nonconjugate mixture of lognormals on survival outcomes. One potential complication arises from the varying dimension of $\pi_1$ depending on the observed atoms in $X_2$. Posterior simulation with variable dimensional parameters generally involves complicated trans-dimensional Markov chain Monte Carlo (Green 1995), often resulting in poor mixing and computational inefficiencies. Our posterior sampling algorithm avoids such complications while rigorously maintaining the architecture of the CAM model. See Section S.8 in the supplementary materials for more details.

6. Simulation Study

We first describe the simulation scenarios.

CAM scenario: We first consider a scenario where the covariates are generated from a CAM model. In this scenario, we take the first $q = p - 3$ covariates to be continuous and the remaining 3 to be binary. For the trial arm $s = 1$, we generate $X_{1,i,1,q} \overset{\text{iid}}{\sim} \sum_{j=1}^{k_{\theta_2}} \pi_{1,j} N_q(\mu_{1,j}, \sigma_j^2 I_1)$ and $X_{1,i,1,\ell} \overset{\text{iid}}{\sim} \text{Bernoulli}(q_1)$ for $\ell = q + 1, \ldots, p$. For the RWD arm, $s = 2$, we generate $X_{2,i,1,q} \overset{\text{iid}}{\sim} \sum_{j=1}^{k_{\theta_2}} \pi_{2,j} N_q(\mu_{2,j}, \sigma_j^2 I_2)$ and $X_{2,i,1,\ell} \overset{\text{iid}}{\sim} \sum_{j=1}^{k_{\theta_2}} \text{Bernoulli}(q_2)$ for $\ell = q + 1, \ldots, p$ where $t_1 = \pi_{2,1} + \pi_{2,2}$ and $t_2 = \pi_{2,3}$. We take $t_1 < t_2$ ensuring that the $X_2$ population is substantially different from $X_1$ in having more heterogeneity.

MIX scenario: In this scenario, we generate $X_{s,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k_{\theta_2}} \pi_{s,j} N_p(\mu_{s,j}, 0.5I_p)$. We take $\mu_{1,j} = \mu_{2,j}$ for all $j < k$ but set $\mu_{1,k} \neq \mu_{2,k}$ so that the atoms in the treatment arm are not exactly a subset of those in the RWD. Given the typically larger heterogeneity of the RWD, this is not a realistic scenario.

We include it to evaluate the approach under model misspecification. Different weights attached to the atoms in the two populations result in significantly different marginal densities.

Interaction scenario: In this scenario, we resample from the historical GBM database of 339 patients to create a future single-arm trial population. Let $F(X)$ denote the (unknown) distribution of the covariates in the database, and $Z$ be an indicator variable such that $Z = s$ if $X$ is selected into arm $s$. That is, we sample $X_{1,i} \text{iid from } p(X_{1,i}) \propto F(X_{1,i}) \cdot c(X_{1,i})$ and $X_{2,i} \text{from } p(X_{2,i}) \propto F(X_{2,i}) \cdot (1 - c(X_{2,i}))$ where $c(X) = \Pr(Z = 1 \mid X)$ is the PS of assignment to the treatment arm. We set $c(X)$ to be a logistic regression with pairwise interactions between some covariates. We can sample $X_{s,i}$ by simple weighted resampling of the historical database, without explicitly knowing $F(.)$.

Oracle scenario: In this fourth and final scenario, we proceed as in the Interaction scenario but now with $c(X)$ defined as a logistic regression with main effects of the true predictors only, that is, as if an oracle had revealed the right predictors.

Outcome model: Under the CAM and MIX scenarios, we generate $Y_{1,i} = \delta + f(X_{1,i}) + \epsilon_{1,i}$ and $Y_{2,i} = f(X_{2,i}) + \epsilon_{2,i}$ where $f(.)$ is a nonlinear function; in the Interaction and Oracle scenarios we generate $Y_{1,i} = \delta + X_{1,i}^T \beta + \epsilon_{1,i}$ and $Y_{2,i} = X_{2,i}^T \beta + \epsilon_{2,i}$ where $\epsilon_{1,i} \sim N(0, 1)$ for $i = 1, \ldots, n_1$ and $s = 1, 2$, implying $\delta$ as the true treatment effect. We repeat the experiments for $\delta = -1, 0, 1, 3$.

We repeat the simulations in the CAM and MIX scenarios for $p = 10, 20, n_1 = 50, 100, 150$ and set $n_2 = 6 \times n_1$ for all setups, keeping the ratio of the population sizes consistent with the GBM application. For each $(n_1, p, \delta)$ combination in the CAM and MIX scenarios, we perform 500 independent replications. Under the Interaction and Oracle scenarios, there are $p = 11$ covariates and we use $n_1 = 49$. To avoid reporting summaries that might just hinge on a lucky choice of the logistic regression coefficients in $c(.)$ and to remove one source of randomness unrelated to the methods under comparison, we independently sample different sets of regression coefficients (from a discrete mixture distribution) for each of the 500 repeat simulations. Further details are provided in Section S.9.2 of the supplementary materials.

6.1. Analyses

We compare the CA-PPMx model with the PS-integrated power prior and composite likelihood approaches (Wang et al. 2019,
as implemented in the \texttt{psrwe} R package, and a two-step population matching approach. We perform seven different analyses for each of the four scenarios to estimate the treatment effect \(\Delta_0\) which we define here as the difference in mean outcomes, that is, \(\delta\). The analyses are (i) \textit{CA-PPMx}: The proposed CA-PPMx model of Section 4.2; (ii) \textit{IS-LM}: The two-step IS approach introduced in Section 3.2. We first sample a subpopulation of size \(n_1\) from \(X_2\) following the importance resampling scheme proposed in Section 3.2 and subsequently estimate the treatment effect between the subpopulation and the treatment arm by fitting a linear model; (iii) and (iv) \textit{PP-Logistic} and \textit{PP-RF}: Two PS-based power prior approaches using logistic regression and random forest (Breiman 2001), respectively; (v) and (vi) \textit{CL-Logistic} and \textit{CL-RF}: Two composite likelihood based approaches with logistic and random forest classifier based PSs, respectively; and finally, (vii) \textit{Matching}: A distance based bipartite matching method designed to match treatment and control groups in observational studies (Hansen and Klopfer 2006) and subsequently using a linear model for detecting treatment effects as implemented in the \texttt{optmatch} R package.

### 6.2. Equivalence of Populations

In preparation for inference under the two-step IS approach, we generate equivalent populations using the density-free importance resampling scheme discussed in Section 3.2 based on the fitted CAM model. To formally test for equivalence of the adjusted datasets, we implement Step 5 in Algorithm 1. We first merge the datasets and then try to classify patients in the merged sample as originally RWD or single-arm treatment cohort \((s = 2\) vs. \(s = 1\) in our earlier notation). For classification, we use BART and report the boxplots of the area under the receiver operating characteristic curve (AUC) of the classification accuracy across the independent experiments for all simulation settings in Figure 3. For comparison, we also subsample randomly (instead of using the IS weights) and report the AUCs in the same figure. We refer to the two sampling strategies as \textit{IS} and \textit{Random}, respectively.

In Figure 3(a), the Random resampling strategy yields high AUC, indicating that the two populations are substantially different and adjustment in the RWD population is necessary before using it as synthetic control. For both, the CAM and MIX scenarios, the performance of the IS scheme improves with increasing sample size. This is expected as for small sample sizes \(X_2\) is lacking enough data to produce a subsample equivalent to \(X_1\). AUC values close to 1 under the CAM scenario imply that the true populations are indeed very different in this case. In contrast, the AUC values close to 0.5 under the IS scheme indicate near equivalence after adjustment. In both scenarios, AUC is substantially reduced under the IS resampling scheme, implying that the proposed CAM model indeed adjusts for the lack of randomization.

Results under the last two scenarios are shown in Figure 3(b). Recall that in both scenarios the simulation truth is not based on the CAM model. Still, the fit under the proposed CAM model achieves near perfect adjustment as shown in the figure.

### 6.3. Inference on Treatment Effects

In each simulation setup, we test \(H_0 : \delta = 0\) versus \(H_1 : \delta \neq 0\) at 5% level of significance. We elaborate the testing procedure in Section S.9.1 of the supplementary materials. We report power in Figure 4, with detailed numerical results appearing in Tables S.1–S.3 in the supplementary materials. Under the PS-based approaches, the power remains below 15% across all scenarios (not shown in the figure). Fully model-based nonparametric
CA-PPMx has higher power than IS-LM and Matching when the true response models are nonlinear. In contrast, the IS-LM and Matching perform comparably and have higher power than the CA-PPMx approach in Interaction and Oracle scenarios where the true response model is linear, but are susceptible to model misspecification as reflected in the CAM and MIX scenarios. This is because IS-LM and Matching assume a linear model for the outcome, which happens to match the simulation truth in the Interaction and Oracle scenarios. Except under the PS-based approaches, power increases with increasing sample size, indicating that PS-based methods may require a much larger population size in the RWD to adjust for the lack of randomization.

7. Application in Glioblastoma

We return to the motivating case study of creating a synthetic control for a hypothetical upcoming single-arm GBM trial. The sample size of the trial is \( n_1 = 49 \), similar to past trials (Vanderbeek et al. 2018). The endpoint of interest is overall survival (OS). We evaluate the operating characteristics of the proposed design by simulating \( L = 100 \) trial replicates. See (Berry et al. 2010, sec. 2.5.4) for a discussion of the role of frequentist operating characteristics in Bayesian inference. To create treatment arm data, we first select covariates \( X_1 \) by randomly selecting patients from the historical database. To generate a realistic nonequivalent patient population, we select not uniformly but using a logistic regression on the covariates (as described in
the Interaction scenario in Section 6). The treatment effect is quantified by the hazard ratio (HR) between the treatment arm and the (synthetic) control arm, with the null and alternative hypotheses $H_0 : HR = 1$ versus $H_1 : HR \leq 0.6$ at 50 weeks. The HR of 0.6 was suggested by clinical collaborators as a meaningful clinical target.

We show results under two alternative scenarios (a) $H_0$: no treatment effect (i.e., $HR = 1$), created by keeping the OS for the patients in the treatment arm as originally observed in the historical database (since the patients received treatments with similar efficacy); and (b) $H_1$: there is a clinically meaningful treatment effect. We created $H_1$ by increasing the OS of patients in the treatment arm with an increment that would correspond to a HR of 0.6 under an exponential model.

We apply three methods to make inference on the treatment effect: (i) IS-based two-step procedure: Here we first create equivalent patient populations using Algorithm 1 and then proceed with inference on the treatment effect as if patients were randomly assigned to treatment and control; (ii) Matching-based two-step procedure: Operationally similar to (i) but now the Matching method discussed in Section 6 is used to create equivalent patient populations; and (iii) Model-based inference: The extension of the CAM model to include the outcomes $Y_{s,i}$, as described in Section 4.2.

### 7.1. IS-based Two-Step Procedure

In preparation for inference, we start with a test for equivalence of the subsampled population in each of the $L = 100$ repeat simulations. Figure 5 plots the relative frequencies for each covariate in the treatment arm (red) and in the synthetic control arm constructed from the RWD using: (a) the IS sampling following Algorithm 1 (green) and (b) random sampling (blue). Very different frequencies in the two arms under random resampling indicate significant differences in the covariate distributions between the treatment and the control arms. For most covariates, the differences are however greatly reduced by the IS scheme.

Once we establish equivalence of the patient populations, we proceed with inference for the treatment effect. We use a Cox proportional hazard (PH) model (Cox 1972) and the logrank test (Peto and Peto 1972) to compare the survival functions. The top panel of Figure 6 shows inference summaries over the $L = 100$ repetitions. The figure shows the histograms of $p$-values under $H_0$ (blue) and $H_1$ (red). Under $H_0$, $p$-values are almost uniformly spread out over $[0, 1]$. In contrast, under $H_1$, the histogram of $p$-values over repeat simulations is peaked close to zero.

Finally, we identify representative simulations from the $L$ repetitions under each of the two scenarios by finding the instance with $p$-value closest to the median of the respective histograms. For these two representatives, we show Kaplan-Meier (KM) survival curves in the top panels of Figure 7(b), respectively. We observe that the survival curves in the two arms are quite alike with wide confidence intervals under the $H_0$ scenario, whereas significant improvements in the survival times can be observed for the treatment arm for the first 80 weeks under the $H_1$ scenario.

### 7.2. Matching-based Two-Step Procedure

We use the Matching procedure to create a synthetic control and then follow the same routine of (i) for inference on treatment
Figure 6. Inference under treatment effects under the two-step procedures: Panel (a) shows histograms of the p-values corresponding to a logrank test under the Cox PH model comparing the survival curves between the treatment arms; Panel (b) shows the Kaplan-Meier curves and pointwise confidence intervals for treatment (blue) and control (red) arms under scenarios $H_0$ (left) and $H_1$, respectively. The top and bottom panels of (a) and (b) show the results corresponding to the IS and Matching based approaches, respectively.

Figure 7. Inference under treatment effects under the model-based approach: Panel (a) shows quantile-quantile plots to assess model fit from Section S.5 in the supplementary materials; the left plot of panel (b) shows posterior probabilities $p(\text{HR} < 0.6 \mid \text{Data})$ under repeat simulations, and on the right posterior estimated hazard ratios for OS with pointwise 95% credible regions are shown under $H_0$ and $H_1$. 
effects. The results are shown in the bottom panels of Figures 6(a) and (b). The distribution of the p-values under the $H_1$ scenario is less peaked around 0 compared to the IS-based procedure. This is also reflected in the representative KM plot under $H_1$ in having a much wider confidence interval around the survival curve possibly indicating the IS-based approach is doing better than Matching in creating equivalent populations.

7.3. Model-Based Inference

As it is not straightforward to account for the uncertainty in creating the synthetic control in the aforementioned two-step procedures, we consider a fully model-based approach. For inference on treatment effects, we first assess goodness of fit procedures, we consider a fully model-based approach. For creating the synthetic control in the aforementioned two-step procedures, we rely on 11 clinically important categorical covariates that are commonly considered as prognostic factors in GBM treatments.

However, in many applications candidate covariates can be high-dimensional. Implicit in the current construction is the assumption that the recorded covariates are clinically relevant for the disease or condition under consideration, and the approach may not be appropriate when large numbers of unscreened candidate covariates are used. Recent advances in Bayesian model-based clustering by Chandra, Canale, and Dunson (2023) could be useful to construct high-dimensional generalizations.

Supplementary Materials

Supplementary materials include additional discussion of the motivating dataset, a brief review on the PPMx, detailed discussion of the graphical goodness-of-fit test for the regression model, an alternative interpretation of our model-based inference approach, choices of hyperparameters, details of the posterior simulation scheme, additional simulation studies and associated details, and MCMC convergence diagnostics. C++ and R programs implementing the methods developed in this article and R Markdown files with instructions are provided in a separately attached Codes.zip folder.

Acknowledgments

We thank the Editor, Dr. Michael Stein, an anonymous Associate Editor, and two anonymous referees for comments that led to significant improvements in the clarity and presentation of the paper.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Noirrit Kiran Chandra @ https://orcid.org/0000-0002-5204-0480

References

Aldape, K., Brindle, K. M., Chesler, L., Chopra, R., Gajjar, A., and Gilbert, M. R. (2019), “Challenges to Curing Primary Brain Tumours,” Nature Reviews Clinical Oncology, 16, 509–520. [2302]

Alexander, B. M., Trippa, L., Gaffey, S., Arrillaga-Romany, I. C., Lee, E. Q., Rinne, M. L., et al. (2019), “Individualized Screening Trial of Innovative Glioblastoma Therapy (INSIGHT): A Bayesian Adaptive Platform Trial to Develop Precision Medicine for Patients with Glioblastoma,” JCO Precision Oncology, 3, 1–13. [2303]

Ascoli, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022), “Clustering Consistency with Dirichlet Process Mixtures,” Biometrika (to appear). [2304]

Au, S., and Beck, J. (2003), “Important Sampling in High Dimensions,” Structural Safety, 25, 139–163. [2305]

Berry, S. M., Carlin, B. P., Lee, J. J., and Müller, P. (2010), Bayesian Adaptive Methods for Clinical Trials, Boca Raton, FL: CRC Press. [2309]

Breiman, L. (2001), “Random Forests,” Machine Learning, 45, 5–32. [2308]

Burcu, M., Dreyer, N. A., Franklin, J. M., Blum, M. D., Critchlow, C. W., Perfetto, E. M., and Zhou, W. (2020), “Real-World Evidence to Support Regulatory Decision-Making for Medicines: Considerations for External Control Arms,” Pharmacoeconomics and Drug Safety, 29, 1228–1235. [2301,2303]

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019), “Latent Nested Nonparametric Priors,” (with discussion), Bayesian Analysis, 14, 1303–1356. [2301,2302]

Chandra, N. K., Canale, A., and Dunson, D. B. (2023), “Escaping the Curse of Dimensionality in Bayesian Model-based Clustering,” Journal of Machine Learning Research, 24, 1–42. [2312]

Chen, M.-H., and Ibrahim, J. G. (2000), “Power Prior Distributions for Regression Models,” Statistical Science, 15, 46–60. [2301]
Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), “Hierarchical Dirichlet Processes,” *Journal of the American Statistical Association*, 101, 1566–1581. [2301,2302,2304]

Vanderbeek, A. M., Rahman, R., Fell, G., Ventz, S., Chen, T., Redd, R., Parmigiani, G., Cloughesy, T. F., Wen, P. Y., Trippa, L., and Alexander, B. M. (2018), “The Clinical Trials Landscape for Glioblastoma: Is It Adequate to Develop New Treatments?” *Neuro-Oncology*, 20, 1034–1043. [2309]

Vansteelandt, S., and Daniel, R. (2014), “On Regression Adjustment for the Propensity Score,” *Statistics in Medicine*, 33, 4053–4072. [2302]

Wang, C., and Rosner, G. L. (2019), “A Bayesian Nonparametric Causal Inference Model for Synthesizing Randomized Clinical Trial and Real-World Evidence,” *Statistics in Medicine*, 38, 2573–2588. [2302]

Wang, C., Li, H., Chen, W.-C., Lu, N., Tiwari, R., Xu, Y., and Yue, L. Q. (2019), “Propensity Score-Integrated Power Prior Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies,” *Journal of Biopharmaceutical Statistics*, 29, 731–748. [2302,2307]

Wang, C., Lu, N., Chen, W.-C., Li, H., Tiwari, R., Xu, Y., and Yue, L. Q. (2020), “Propensity Score-Integrated Composite Likelihood Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies,” *Journal of Biopharmaceutical Statistics*, 30, 495–507. [2308]

Zhao, Z. (2004), “Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence,” *The Review of Economics and Statistics*, 86, 91–107. [2302]