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

Noirrit Kiran Chandra[a] (noirrit.chandra@utdallas.edu)
Abhra Sarkar[b] (abhra.sarkar@utexas.edu)
John F. de Groot[c] (john.degroot@ucsf.edu)
Ying Yuan[d] (yyuan@mdanderson.org)
Peter Müller[b,e] (pmueller@math.utexas.edu)

[a]Department of Mathematical Sciences,
The University of Texas at Dallas, TX, USA

[b]Department of Statistics and Data Sciences,
The University of Texas at Austin, TX, USA

[c]Department of Neurological Surgery,
University of California San Francisco, CA, USA

[d]Department of Biostatistics,
The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[e]Department of Mathematics,
The University of Texas at Austin, TX, USA

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 paper, 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 non-linear response functions. We apply the method to supplement single arm treatment-only glioblastoma studies with a synthetic control arm based on historical trials.

Key Words: common atoms mixture, glioblastoma, importance sampling, mixtures, real world data, single-arm trials.

Short/Running Title: Common Atoms Mixture Model for Synthetic Controls

Corresponding Author: Noirrit Kiran Chandra (noirrit.chandra@utdallas.edu)

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); Rodríguez *et al.* (2008); 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 *et al.*, 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 *et al.*, 2000; Chen and Ibrahim, 2000), commensurate prior (Hobbs *et al.*, 2011) and elastic prior (Jiang *et al.*, 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 *et al.* (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 sen-

sitive 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 (Vansteelandt and Daniel, 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 data set 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 (Rodríguez *et al.*, 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 paper 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 paper for easy reference.

Table 1: List of acronyms

Acronym	Full forms	Acronym	Full forms
AUC	area under the receiver operating characteristic curve	DP	Dirichlet process
		GBM	glioblastoma
BART	Bayesian additive regression tree	IS	importance sampling
		PPMx	product partition model with regression on covariates
BNP	Bayesian nonparametric		
CAM	common atoms mixture	PS	propensity score
CA-PPMx	common atoms PPMx	RCT	randomized controlled trial
		RWD	real-world data

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 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 towards 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.

Table 2: Description of the covariates in the GBM data.

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 (concurrent radiation therapy and temozolomide): Yes/No
CT	participation in a therapeutic trial: Yes/No
MGMT	status of MGMT (O^6-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)

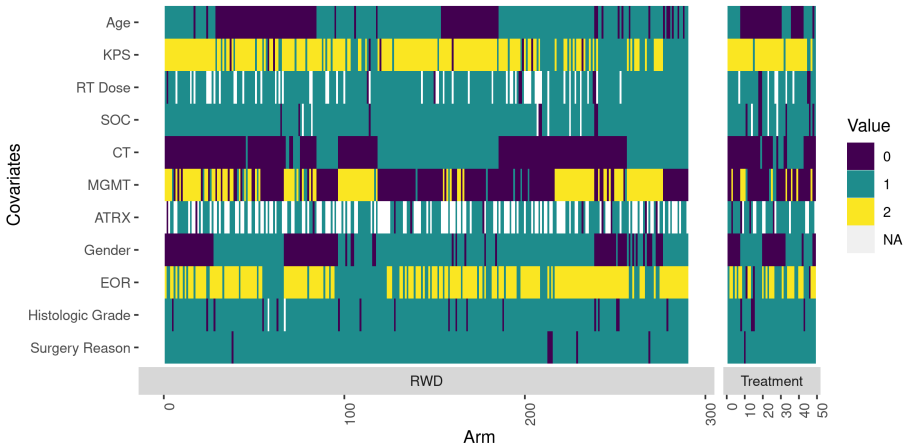


Figure 1: Glioblastoma dataset of 11 baseline categorical covariates with missing entries in the two treatment arms. The left block shows the historical patients. The (smaller) right block shows a hypothetical future trial.

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 $(\boldsymbol{X}_{s,i}, Y_{s,i})$, $s = 1, 2, \ldots, S$, comprising p-dimensional covariate vectors $\boldsymbol{X}_{s,i} = (X_{s,i,1}, \ldots, X_{s,i,p})^{\mathrm{T}}$ and corresponding responses $Y_{s,i}$ 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, i.e., when $S > 2$, we would simply merge them and consider the merged data set 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, i.e., matching the distributions of $\boldsymbol{X}_{s,i}$ 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, 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 $\boldsymbol{X}_{1,i}$ and a matching partition of $\boldsymbol{X}_{2,i}$. 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 $\boldsymbol{X}_{2,i}$ in the RWD. Let $\widetilde{\boldsymbol{\zeta}} = \{\widetilde{\boldsymbol{\zeta}}_j\}_{j=1}^{\infty}$ and $\boldsymbol{\pi}_2 = \{\pi_{2,j}\}_{j=1}^{\infty}$ denote cluster-specific parameters and weights, respectively. We let

$$\boldsymbol{X}_{2,i} \mid \widetilde{\boldsymbol{\zeta}}, \boldsymbol{\pi}_2 \overset{\text{iid}}{\sim} \overbrace{\textstyle\sum_{j=1}^{\infty} \pi_{2,j} q(\boldsymbol{X}_{2,i} \mid \widetilde{\boldsymbol{\zeta}}_j)}^{F_2(\boldsymbol{X}_{2,i} \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}})}, \quad \widetilde{\boldsymbol{\zeta}}_j \mid \boldsymbol{\xi} \overset{\text{iid}}{\sim} G_0(\widetilde{\boldsymbol{\zeta}}_j \mid \boldsymbol{\xi}), \quad \boldsymbol{\pi}_2 \sim \mathrm{GEM}(\alpha_2). \tag{1}$$

Here $q(\cdot \mid \widetilde{\boldsymbol{\zeta}}_j)$ is a suitably chosen kernel with parameter $\widetilde{\boldsymbol{\zeta}}_j$, $G_0(\cdot \mid \boldsymbol{\xi})$ is a prior distribution for the $\widetilde{\boldsymbol{\zeta}}_j$'s, and $\mathrm{GEM}(\alpha)$ is a stick-breaking prior on the mixture weights corresponding to a DP with mass parameter $\alpha > 0$ (Sethuraman, 1994). Let $G =$

$\sum_{j=1}^{\infty} \pi_{2,j}\delta_{\widetilde{\boldsymbol{\zeta}}_j}(\cdot)$ denote a discrete probability measure with atoms at the $\widetilde{\boldsymbol{\zeta}}_j$'s. An equivalent hierarchical model representation of (1) is

$$\boldsymbol{X}_{2,i} \mid \boldsymbol{\zeta}_i \overset{\text{iid}}{\sim} q(\boldsymbol{X}_{2,i} \mid \boldsymbol{\zeta}_i), \quad \boldsymbol{\zeta}_i \mid G \overset{\text{iid}}{\sim} G, \quad G \mid \alpha_2, \boldsymbol{\xi} \sim \text{DP}\{\alpha_2, G_0(\cdot \mid \boldsymbol{\xi})\}, \tag{2}$$

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

$$\boldsymbol{X}_{1,i} \mid \boldsymbol{\zeta}^\star \overset{\text{iid}}{\sim} \overbrace{\textstyle\sum_{j=1}^{k(n_2)} \pi_{1,j} q(\boldsymbol{X}_{1,i} \mid \boldsymbol{\zeta}_j^\star)}^{F_1(\boldsymbol{X}_{1,i} \mid \boldsymbol{\pi}_1, \boldsymbol{\zeta}^\star)}, \quad \boldsymbol{\pi}_1 \sim \text{Dir}\left\{\tfrac{\alpha_1}{k(n_2)}, \ldots, \tfrac{\alpha_1}{k(n_2)}\right\}, \tag{3}$$

where $\boldsymbol{\pi}_1 = (\pi_{1,1}, \ldots, \pi_{1,k(n_2)})$, $\text{Dir}(a_1, \ldots, a_r)$ indicates an r-dimensional Dirichlet distribution with parameters $a_1, \ldots, a_r$, and $\alpha_1 > 0$ is a concentration parameter. Note that model (3) is defined conditionally on (1) and $\boldsymbol{\zeta}^\star$ such that F_1 and F_2 share the same set of atoms. Importantly, the construction avoids the imputation of clusters (strata) with only $\boldsymbol{X}_{1,i}$'s. There is always a corresponding (non-empty) cluster for the $\boldsymbol{X}_{2,i}$'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, $\boldsymbol{X}_2$ can be expected to exhibit greater heterogeneity than $\boldsymbol{X}_1$ (see, e.g., the right panel in Figure 2).

In summary, we define $F_1(\boldsymbol{X} \mid \boldsymbol{\pi}_1, \boldsymbol{\zeta}^\star) = \sum_{j=1}^{k(n_2)} \pi_{1,j} q(\boldsymbol{X} \mid \boldsymbol{\zeta}_j^\star)$ and $F_2(\boldsymbol{X} \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}}) = \sum_{j=1}^{\infty} \pi_{2,j} q(\boldsymbol{X} \mid \widetilde{\boldsymbol{\zeta}}_j)$, 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 $\widetilde{\boldsymbol{\zeta}}_3$ is not linked with any $\boldsymbol{X}_2$ observation and hence $k(n_2) = 3$.

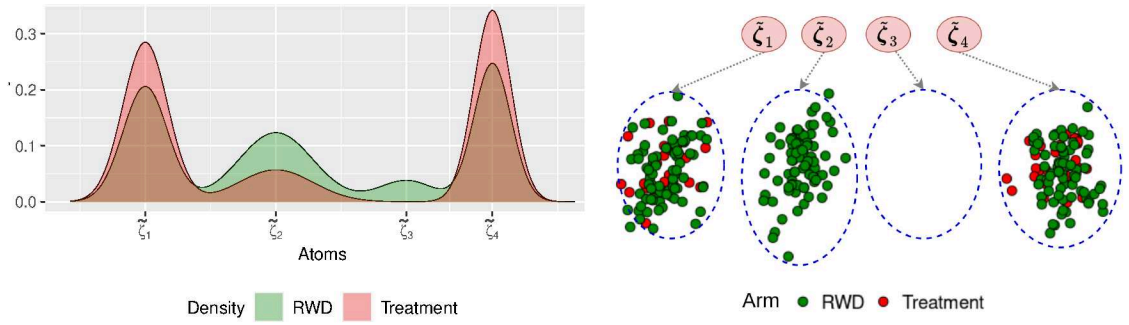


Figure 2: An illustration of the CAM model: In the generative model, there are a total of four atoms $\widetilde{\boldsymbol{\zeta}}_{1:4}$ shared between RWD and the treatment arm (left panel). Despite having positive weight, the atom $\widetilde{\boldsymbol{\zeta}}_3$ is not associated with any sample from the RWD (right panel) and hence the density of the treatment arm is also allowed to be supported only on the remaining non-empty clusters $\widetilde{\boldsymbol{\zeta}}_1, \widetilde{\boldsymbol{\zeta}}_2, \widetilde{\boldsymbol{\zeta}}_4$ of the RWD. The atom $\widetilde{\boldsymbol{\zeta}}_2$ is linked with only the RWD (right panel). A cluster for the treatment arm alone is however not permissible.

Accordingly, $F_1(\cdot \mid \boldsymbol{\pi}_1, \boldsymbol{\zeta}^\star)$ is a mixture of three components. Finally, no observation from $\boldsymbol{X}_1$ is linked to $\widetilde{\boldsymbol{\zeta}}_2$. The $\boldsymbol{X}_{2,i}$'s linked to $\widetilde{\boldsymbol{\zeta}}_1$ and $\widetilde{\boldsymbol{\zeta}}_4$ can be regarded as digital clones of the $\boldsymbol{X}_{1,i}$'s linked to the same atoms.

The described CAM model is different from existing BNP mixture models. In (1)-(3), the atoms linked to $\boldsymbol{X}_1$ are always a subset of those atoms that are linked to $\boldsymbol{X}_2$, which is not naturally the case for the hierarchical DP model (Teh *et al.*, 2006). Also, unlike the nested DP (Rodríguez *et al.*, 2008) and the common atoms nested DP (Denti *et al.*, 2021) models, there is no notion of clustering distributions. That is, $p\{F_1(\boldsymbol{X} \mid \boldsymbol{\pi}_1, \boldsymbol{\zeta}^\star) = F_2(\boldsymbol{X} \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}})\} = 0$ a priori. Instead, the intention here is to cluster similar covariate values across the datasets.

Regarding the concentration parameters α_s, we assume $\log \alpha_s \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$ for $s = 1, 2$. 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).

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 *et al.*, 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, Section 7.2.2, pp 175). Specifically, let $\mathcal{O}_{s,i} = \{j : X_{s,i,j} \text{ is recorded}\}$ denotes the set of observed covariates for patient i in dataset s. We use independent kernels

$$q(\boldsymbol{X}_{s,i} \mid \boldsymbol{\zeta}_j^\star) = \prod_{\ell \in \mathcal{O}_{s,i}} q_\ell(X_{s,i,\ell} \mid \boldsymbol{\zeta}_{j,\ell}^\star), \quad G_0(\boldsymbol{\zeta}_j^\star \mid \boldsymbol{\xi}) = \prod_{\ell=1}^{p} g_{0,\ell}(\boldsymbol{\zeta}_{j,\ell}^\star \mid \boldsymbol{\xi}), \tag{4}$$

where $q_\ell(\cdot \mid \boldsymbol{\zeta}_{j,\ell}^\star)$ is a univariate kernel corresponding to the ℓ^{th} covariate with parameters $\boldsymbol{\zeta}_{j,\ell}^\star$ and $g_{0,\ell}(\boldsymbol{\zeta}_{j,\ell}^\star \mid \boldsymbol{\xi})$ is a prior on $\boldsymbol{\zeta}_{j,\ell}^\star$ with hyper-parameters $\boldsymbol{\xi}$. The likelihood function of $\boldsymbol{X}_{s,i}$ is then computed on the basis of only the observed values. The kernel q_ℓ is chosen to accommodate the data-type of the ℓ^{th} covariate. The model allows co-clustering of $\boldsymbol{X}_{s,i}$ with some missing variables and another fully observed $\boldsymbol{X}_{s,i'}$; 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, e.g., 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 $\boldsymbol{X}_2$ that can be considered to be equivalent to $\boldsymbol{X}_1$ (see below for a definition of equivalence that is being used here). Under the assumption of no unmeasured confounders, the $\boldsymbol{X}_{2,i}$'s in the sampled (or weighted) subpopulation can be assumed to follow the same distribution as $\boldsymbol{X}_{1,i}$, and be considered digital clones of the $\boldsymbol{X}_{1,i}$. 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_s denotes the mixture model for $\boldsymbol{X}_{s,i}$, $s = 1, 2$, under (1) and (3), respectively. We define equivalent populations as a subset (possibly all) of $\boldsymbol{X}_2$ together with a set of weights such that expectation of any function of interest $g(\boldsymbol{X}_{1,i})$ under $\boldsymbol{X}_{1,i} \sim F_1$ can be evaluated as a (weighted) Monte Carlo average using these $\boldsymbol{X}_2$ (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 $F_2(\cdot \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}}) = \sum_{j=1}^{\infty} \pi_{2,j} q(\cdot \mid \widetilde{\boldsymbol{\zeta}}_j)$. Alternatively, the joint model of $(\boldsymbol{X}_2, \boldsymbol{c}_2)$ can be expressed as $F_2^H(\boldsymbol{X}_{2,i}, c_{2,i} \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}}) = q(\boldsymbol{X}_{2,i} \mid \widetilde{\boldsymbol{\zeta}}_{c_{2,i}}) \pi_{2,c_{2,i}}$. For easier housekeeping, we assume $\boldsymbol{\zeta}_j^\star = \widetilde{\boldsymbol{\zeta}}_j$ for $j = 1, \ldots, k(n_2)$, i.e., the first $k(n_2)$ atoms are linked with the $\boldsymbol{X}_{2,i}$'s. Accordingly, we let $F_1(\cdot \mid \boldsymbol{\pi}_1, \widetilde{\boldsymbol{\zeta}}) = \sum_{j=1}^{k(n_2)} \pi_{1,j} q(\cdot \mid \widetilde{\boldsymbol{\zeta}}_j)$ using the same first $k(n_2)$ atoms observed in the $\boldsymbol{X}_2$ population. This is the exact construction of (1) and (3). For an equivalent population, we require weights w_i attached to $(\boldsymbol{X}_{2,i}, c_{2,i})$ (using $w_i = 0$ to drop samples) such that:

$$\mathbb{E}_{F_1(\cdot \mid \boldsymbol{\pi}_1, \widetilde{\boldsymbol{\zeta}})} \{g(\boldsymbol{X}_{1,i})\} = \mathbb{E}_{F_2^H(\cdot \mid \boldsymbol{\pi}_2, \widetilde{\boldsymbol{\zeta}})} \{\widehat{g}(\boldsymbol{X}_2, \boldsymbol{c}_2)\} \text{ with } \widehat{g}(\boldsymbol{X}_2, \boldsymbol{c}_2) = \sum_{i=1}^{n_2} w_i \, g(\boldsymbol{X}_{2,i}, c_{2,i}).$$

The weights w_i are functions of $c_{2,i}$ and $\pi_{1,j}$ as follows. Define $n_{2,j} = |C_{2,j}|$, the cardi-

nality of the earlier introduced clusters $C_{2,j}$. Then $\frac{1}{n_{2,j}} \sum_{i \in C_{2,j}} g(\boldsymbol{X}_{2,i})$ is an unbiased estimator of $\mathbb{E}_{q(\cdot|\widetilde{\boldsymbol{\zeta}}_j)} g(\boldsymbol{X})$ and

$$\widehat{g} = \sum_j \pi_{1,j} \left\{ \sum_{i \in C_{2,j}} \frac{1}{n_{2,j}} g(\boldsymbol{X}_{2,i}) \right\} = \sum_{i=1}^{n_2} \frac{\pi_{1,c_{2,i}}}{n_{2,c_{2,i}}} g(\boldsymbol{X}_{2,i}) \tag{5}$$

is an unbiased estimator of $\mathbb{E}_{F_1(\cdot|\boldsymbol{\pi}_1, \widetilde{\boldsymbol{\zeta}})} g(\boldsymbol{X})$. We then recognize $\pi_{1,c_{2,i}}/n_{2,c_{2,i}}$ as the ideal weights. Since we only observe $\boldsymbol{X}_s$ but not $\boldsymbol{\pi}_1$ and $\boldsymbol{c}_2$, we replace $\pi_{1,c_{2,i}}/n_{2,c_{2,i}}$ in $\widehat{g}$ by a Monte Carlo average under 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 $\pi_{1,j}^{(m)}$, $n_{2,j}^{(m)}$, etc. to indicate parameter values in the m^{th} sample. We use

$$\widehat{g} = \sum_i w_i g(\boldsymbol{X}_{2,i}), \quad w_i \propto \sum_{m=1}^{M} \pi^{(m)}_{1,c^{(m)}_{2,i}} / n^{(m)}_{2,c^{(m)}_{2,i}}, \tag{6}$$

with w_i being the *importance sampling* weight for $\boldsymbol{X}_{2,i}$. The $\boldsymbol{X}_{2,i}$'s can be resampled with these weights to obtain the desired subpopulation with distribution $F_1(\boldsymbol{x} \mid \boldsymbol{\pi}_1, \widetilde{\boldsymbol{\zeta}})$ (Skare *et al.*, 2003). This resampled subpopulation of $\boldsymbol{X}_2$ can then be regarded as equivalent in distribution to $\boldsymbol{X}_1$. Algorithm 1 summarizes the procedure.

Algorithm 1: Density-free importance resampling of RWD and validation

1. **Input** two data sets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.
2. **Fit the CAM model** to the data using MCMC simulation. Let $\boldsymbol{\pi}_1^{(m)}$ and $\boldsymbol{c}_2^{(m)}$ be the m^{th} MCMC sample of $\boldsymbol{\pi}_1$ and $\boldsymbol{c}_2$, respectively, and $n_{2,j}^{(m)}$ be the size of cluster $C_{2,j}$ in the m^{th} MCMC iteration for $m = 1, \ldots, M$.
3. **Calculate** importance sampling weights
$$w_i \propto \sum_{m=1}^{M} \pi^{(m)}_{1,c^{(m)}_{2,i}} / n^{(m)}_{2,c^{(m)}_{2,i}}, \quad i = 1, \ldots, n_2.$$
4. **Resample** a subpopulation of size n_1 from $\boldsymbol{X}_2$ with importance resampling weights w_i with replacement.
5. **Test for equivalence** of $\boldsymbol{X}_1$ and the resampled subpopulation of $\boldsymbol{X}_2$ using a supervised classification algorithm (e.g., a BART as described in the text).

To test the equivalence of the two populations, we use a Bayesian additive regression tree (BART, Chipman *et al.*, 2010) in Step 5 of Algorithm 1. In extensive simulation studies in Section 6, we notice that an AUC (area under the receiver oper-

ating 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 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_1) 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 ≥ 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 $\boldsymbol{Y}$. 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 $Y_{s,i}$ on covariates $\boldsymbol{X}_{s,i}$ by first grouping patients with similar covariate profiles into clusters and then adding a cluster-specific sampling model for the outcome $Y_{s,i}$. That is, the overall model specifies a regression of $Y_{s,i}$ on $\boldsymbol{X}_{s,i}$ via a random partition. A major advantage of this approach is that it allows a variable-dimension covariate vector – a feature that is not straightforward to include in a regression otherwise. Similar product partition models with regression on covariates (PPMx, see also S.2 in the supplementary materials) were considered by Müller *et al.* (2011) and Page *et al.* (2022), albeit without any notion of common atoms. We will therefore refer to the model proposed below as the common atoms PPMx (CA-PPMx). Formally, we introduce cluster-specific parameters $\boldsymbol{\theta}_s = \{\boldsymbol{\theta}_{s,j};\ j = 1, \ldots, k(n_2)\}$, and assume

$$(Y_{s,i} \mid \boldsymbol{\theta}_s, c_{s,i} = j) \overset{\text{ind}}{\sim} h(Y_{s,i} \mid \boldsymbol{\theta}_{s,j}), \tag{7}$$

for a suitable choice of h. For example, for an event-time response, h could be a lognormal, exponential or Weibull model. The response model (7) depends on the covariates indirectly via $c_{s,i}$'s, i.e., the partition induced by the covariates. Within stratum $C_j = C_{1,j} \cup C_{2,j}$, the response models allows for a treatment comparison based on $(\boldsymbol{\theta}_{1,j}, \boldsymbol{\theta}_{2,j})$, which can then be averaged with respect to the assumed distribution of $\boldsymbol{X}_{s,j}$ to define an average treatment effect.

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

$$Y_{s,i} \mid c_{s,i} = j, \boldsymbol{\theta}_s \overset{\text{ind}}{\sim} \text{N}(\mu_{s,j}, \sigma^2_{s,j}),\ \ \mu_{s,j} \mid \sigma^2_{s,j} \overset{\text{ind}}{\sim} \text{N}\left(\mu_0, \tfrac{\sigma^2_{s,j}}{\kappa_0}\right),\ \ \sigma^{-2}_{s,j} \overset{\text{iid}}{\sim} \text{Ga}(a_0, b_0), \tag{8}$$

where $\text{Ga}(a_0, b_0)$ is a gamma distribution with mean a_0/b_0. We add the hyper-priors $\mu_0 \sim \text{N}(m_\mu, s_\mu^2)$ and $\log b_0 \sim \text{N}(m_b, s_b^2)$ on the main location-scale controlling hyper-parameters μ_0 and b_0 while fixing the precision hyper-parameters κ_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 Militký, 2011, Chapter 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 $\boldsymbol{X}_2$, 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, i.e., 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, $\Delta_\theta = \delta\left\{f_1(\cdot \mid \boldsymbol{\theta}_1, \boldsymbol{\pi}_1), f_2(\cdot \mid \boldsymbol{\theta}_2, \boldsymbol{\pi}_2)\right\}$. 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 $f_1(\boldsymbol{Y}_{1,i} \mid \boldsymbol{\theta}_1, \boldsymbol{\pi}_1)$ and $f_2(\boldsymbol{Y}_{2,i} \mid \boldsymbol{\theta}_2, \boldsymbol{\pi}_2)$ 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

$$\widetilde{f}_2(Y \mid \boldsymbol{\theta}_2, \boldsymbol{\pi}_1) = \textstyle\sum_{j=1}^{k(n_2)} \pi_{1,j} h(Y \mid \boldsymbol{\theta}_{2,j}),$$

where the mixture components $h(Y \mid \boldsymbol{\theta}_{2,j})$ of the response model in the RWD are weighted by $\boldsymbol{\pi}_1$, i.e., the cluster weights associated with $\boldsymbol{X}_1$ (rather than $\boldsymbol{\pi}_2$). Thus $\widetilde{f}_2$ 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

$$\widetilde{\Delta}_\theta = \delta\{f_1(\cdot \mid \boldsymbol{\theta}_1, \boldsymbol{\pi}_1), \widetilde{f}_2(\cdot \mid \boldsymbol{\theta}_2, \boldsymbol{\pi}_1)\}. \tag{9}$$

For example, when Y is a univariate response variable and $\delta(f_1, f_2) = \mathbb{E}_{f_1}(Y) - \mathbb{E}_{f_2}(Y)$, $\widetilde{\Delta}_\theta$ simplifies to $\widetilde{\Delta}_\theta = \sum_{j=1}^{k(n_2)} \pi_{1,j} \left\{\mathbb{E}_{h(Y|\boldsymbol{\theta}_{1,j})}(Y) - \mathbb{E}_{h(Y|\boldsymbol{\theta}_{2,j})}(Y)\right\}$, which further reduces to $\widetilde{\Delta}_\theta = \sum_{j=1}^{k(n_2)} \pi_{1,j}(\mu_{1,j} - \mu_{2,j})$ when $\mu_{s,j} = \mathbb{E}_{h(Y|\boldsymbol{\theta}_{s,j})}\{T(Y)\}$.

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 $\delta\{h(Y \mid \boldsymbol{\theta}_{1,j}), h(Y \mid \boldsymbol{\theta}_{2,j})\}$ 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 non-conjugate mixture of lognormals on survival outcomes. One potential complication arises from the varying dimension of $\boldsymbol{\pi}_1$ depending on the observed atoms in $\boldsymbol{X}_2$. 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 $\boldsymbol{X}_{1,i,1:q} \overset{\text{iid}}{\sim} \sum_{j=1}^{2} \pi_{1,j} \mathrm{N}_q(\boldsymbol{\mu}_j, \sigma_j^2 \boldsymbol{I}_q)$ and $X_{1,i,\ell} \overset{\text{iid}}{\sim} \text{Bernoulli}(\varrho_1)$ for $\ell = q{+}1, \ldots, p$. For the RWD arm, $s = 2$, we generate $\boldsymbol{X}_{2,i,1:q} \overset{\text{iid}}{\sim} \sum_{j=1}^{3} \pi_{2,j} \mathrm{N}_q(\boldsymbol{\mu}_j, \sigma_j^2 \boldsymbol{I}_q)$ and $X_{2,i,\ell} \overset{\text{iid}}{\sim} \sum_{j=1}^{2} \iota_j \text{Bernoulli}(\varrho_j)$ for $\ell = q{+}1, \ldots, p$ where $\iota_1 = \pi_{2,1} + \pi_{2,2}$ and $\iota_2 = \pi_{2,3}$. We take $\iota_1 \ll \iota_2$ ensuring that the $\boldsymbol{X}_2$ population is substantially different from $\boldsymbol{X}_1$

in having more heterogeneity.

MIX scenario: In this scenario, we generate $\boldsymbol{X}_{s,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \pi_{s,j} \mathrm{N}_p(\boldsymbol{\mu}_{s,j}, 0.05\boldsymbol{I}_p)$. We take $\boldsymbol{\mu}_{1,j} = \boldsymbol{\mu}_{2,j}$ for all $j < k$ but set $\boldsymbol{\mu}_{1,k} \neq \boldsymbol{\mu}_{2,k}$ 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 $F(\boldsymbol{X})$ denote the (unknown) distribution of the covariates in the database, and Z be an indicator variable such that $Z = s$ if $\boldsymbol{X}$ is selected into arm s. That is, we sample $\boldsymbol{X}_{1,i}$ i.i.d. from $p(\boldsymbol{X}_{1,i}) \propto F(\boldsymbol{X}_{1,i}) \cdot e(\boldsymbol{X}_{1,i})$ and $\boldsymbol{X}_{2,i}$ from $p(\boldsymbol{X}_{2,i}) \propto F(\boldsymbol{X}_{2,i}) \cdot \{1 - e(\boldsymbol{X}_{2,i})\}$ where $e(\boldsymbol{X}) = \Pr(Z = 1 \mid \boldsymbol{X})$ is the PS of assignment to the treatment arm. We set $e(\boldsymbol{X})$ to be a logistic regression with pairwise interactions between some covariates. We can sample $\boldsymbol{X}_{s,i}$ by simple weighted resampling of the historical database, without explicitly knowing $F(\cdot)$.

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

Outcome model: Under the CAM and MIX scenarios, we generate $Y_{1,i} = \delta + f(\boldsymbol{X}_{1,i}) + \epsilon_{1,i}$ and $Y_{2,i} = f(\boldsymbol{X}_{2,i}) + \epsilon_{2,i}$ where $f(\cdot)$ is a nonlinear function; in the Interaction and Oracle scenarios we generate $Y_{1,i} = \delta + \boldsymbol{X}_{1,i}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{1,i}$ and $Y_{2,i} = \boldsymbol{X}_{2,i}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{2,i}$, where $\epsilon_{s,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$ for $i = 1, \ldots, n_s$ and $s = 1, 2$, implying δ 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, δ) 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 $e(\cdot)$ 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.

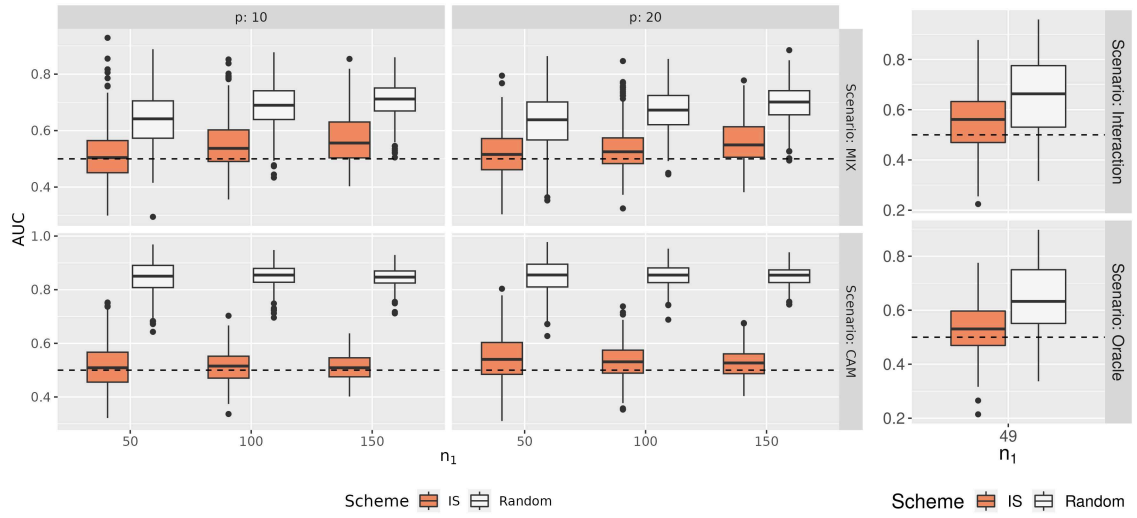
Analyses: We compare the CA-PPMx model with the PS-integrated power prior and composite likelihood approaches (Wang *et al.*, 2019, 2020; Chen *et al.*, 2020) as implemented in the `psrwe` `R` package, and a two-step population *matching* approach. We perform seven different analyses for each of the four scenarios to estimate the treatment effect Δ_θ which we define here as the difference in mean outcomes, i.e., δ. The analyses are (i) *CA-PPMx*: The proposed *CA-PPMx* model of Section 4.2; (ii) *IS-LM*: The two-step IS approach introduced in Section 3.2. We first sample a subpopulation of size n_1 from $\boldsymbol{X}_2$ 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) *PP-Logistic* and *PP-RF*: Two PS-based power prior approaches using logistic regression and random forest (Breiman, 2001), respectively; (v) and (vi) *CL-Logistic* and *CL-RF*: Two composite likelihood based approaches with logistic and random forest classifier based PSs, respectively; and finally, (vii) *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 `optmatch R` package.

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 *IS* and *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 $\boldsymbol{X}_2$ is lacking enough data to produce a subsample equivalent to $\boldsymbol{X}_1$. 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.

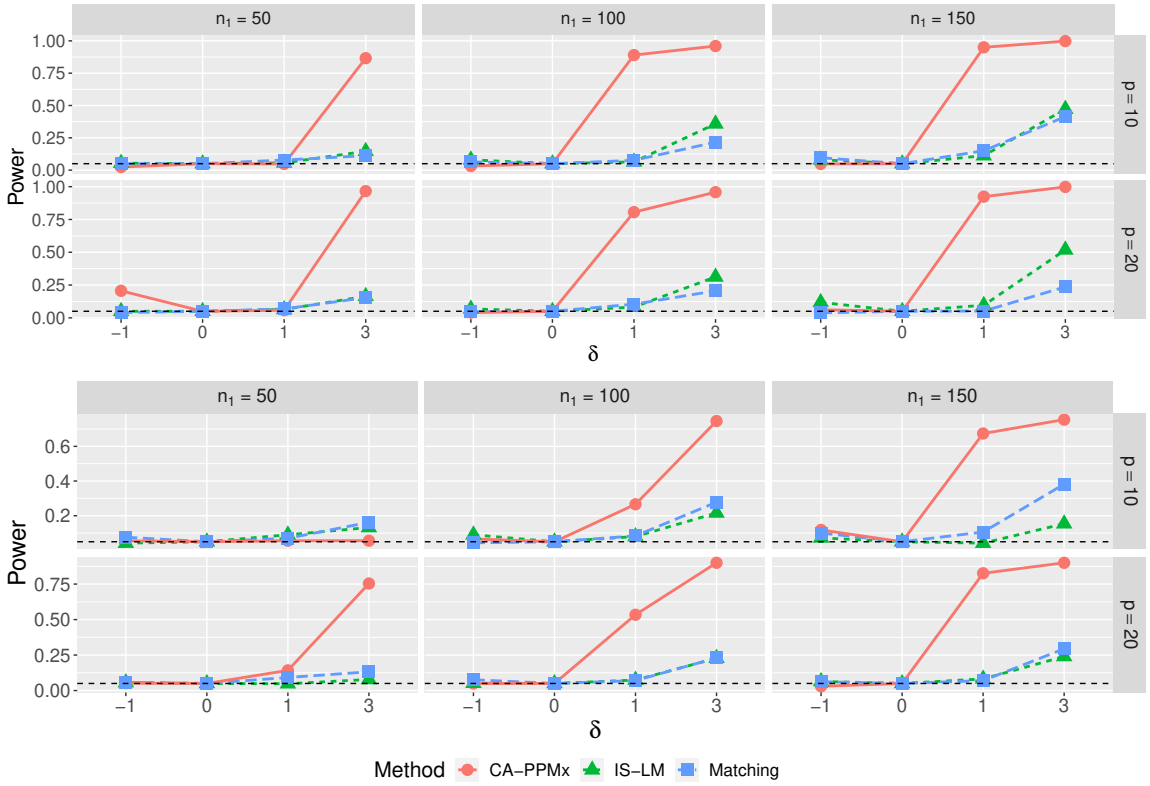
Results under the last two scenarios are shown in Figure 3(b). Recall that in both scenarios the simulation truth is not based on the CAM model. Still, the fit under the proposed CAM model achieves near perfect adjustment as shown in the figure.



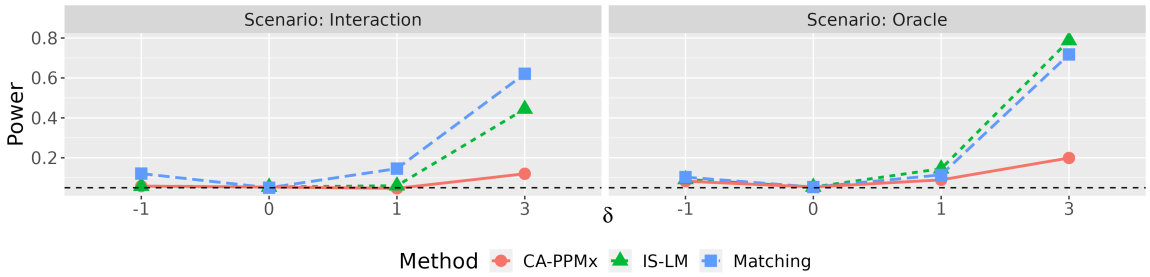
(a) (b)

Figure 3: Boxplots of the area under the receiver operating characteristic curve (AUC) of the classification accuracy for a merged dataset consisting of $\boldsymbol{X}_1$ and the subsampled $\boldsymbol{X}_2$ using BART, and trying to classify into originally $\boldsymbol{X}_1$ versus $\boldsymbol{X}_2$. Two subsampling schemes are used - the importance sampling (IS) strategy in Section 3.2 and simple random resampling. Here AUCs close to 0.5 imply near equivalence between the populations. Panel (a) shows the AUCs in the CAM and MIX scenarios across different sample sizes and number of covariates; and (b) shows the AUCs under the Interaction and Oracle scenarios.

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.2 and 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 non-linear. 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



(a) CAM (top) and MIX (bottom) scenarios.



(b) Interaction (left) and oracle (right) scenarios.

Figure 4: Power of detecting treatment effects in different simulation setups for a 5% level of significance test: Seven methods are used to estimate the effects where IS-LM and CA-PPMx are based on the proposed CAM model. Panel (a) corresponds to the CAM (top) and MIX (bottom) scenarios. Panel (b) shows results under the Interaction (left side) and the Oracle (right side) scenarios.

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, Section 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 $\boldsymbol{X}_{1,i}$ by randomly selecting patients from the historical database. To generate a realistic non-equivalent 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 $H_0 : \text{HR} = 1$ vs. $H_1 : \text{HR} \leq 0.6$ 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.

(i) 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.

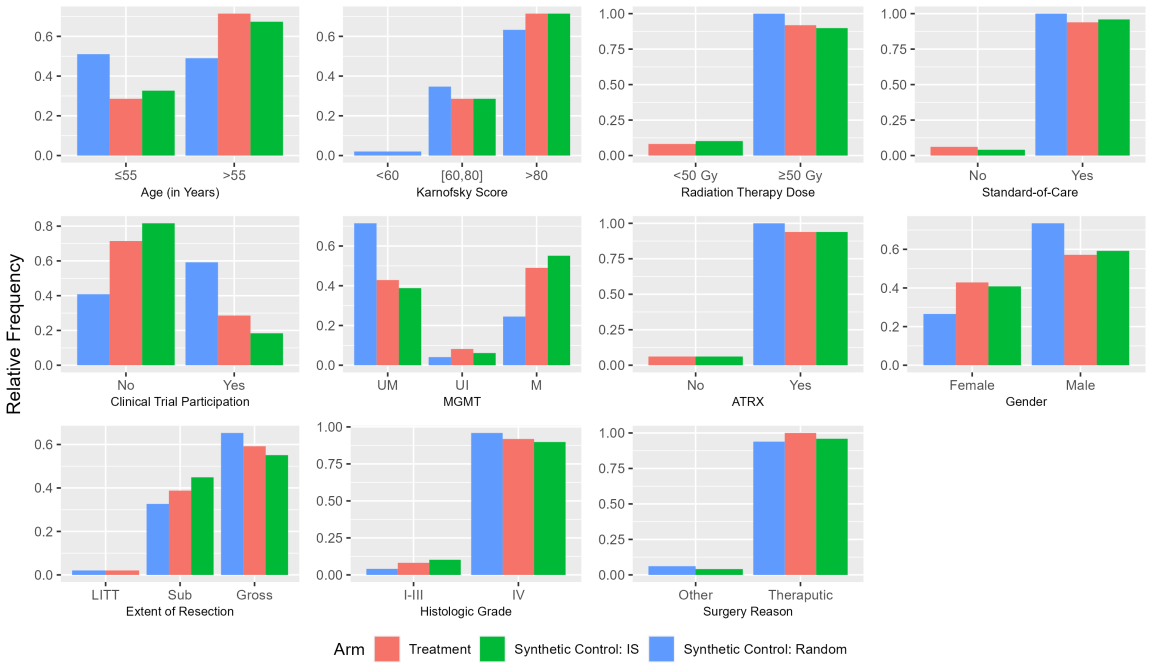
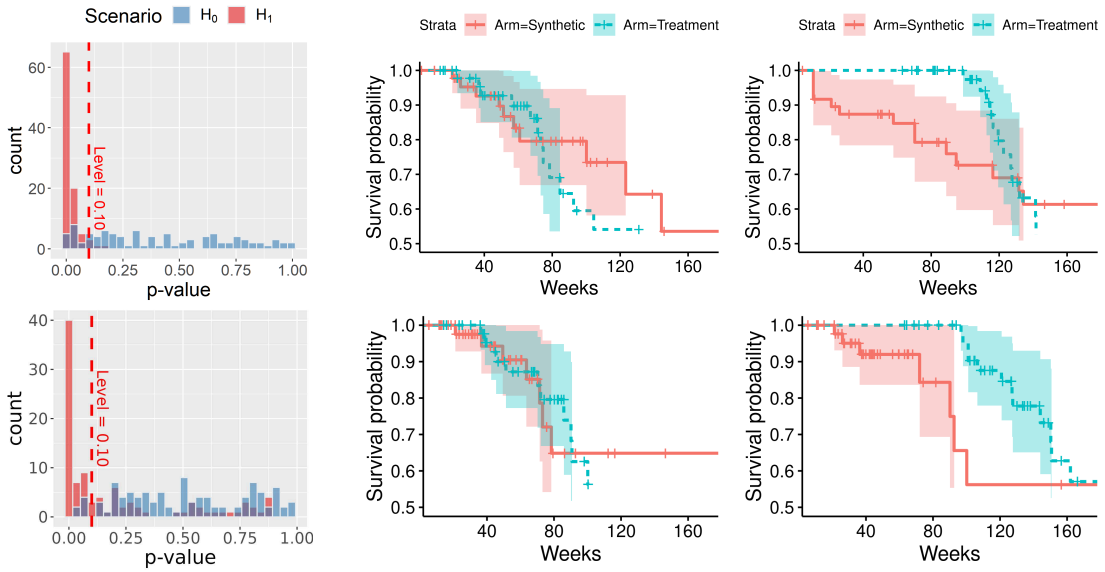


Figure 5: Covariate distributions before and afte adjustments. The red bars show the distributions of the covariates in the treatment arm. The green and blue bars show the distributions of the covariates in the synthetic control arms formed using the IS and random resampling schemes, respectively.



(a) p-values under the Cox PH model.

(b) Kaplan-Meier curves under H_0 (left) and H_1 (right) scenarios.

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.

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(a) 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.

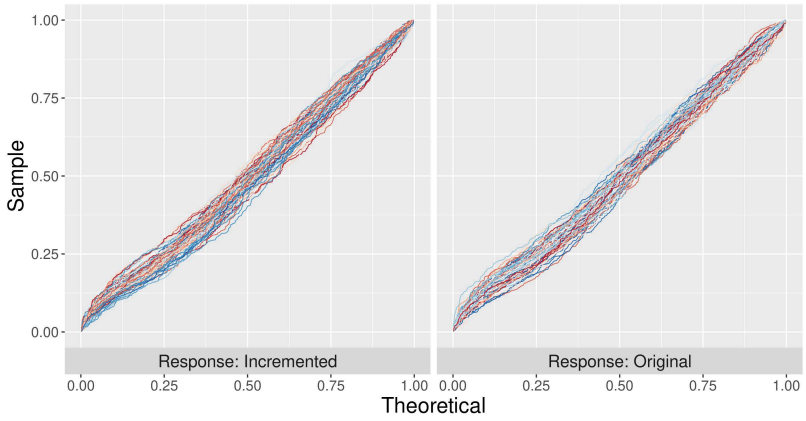
Finally, we identify representative simulations from the L repetitions under each of the two scenarios by finding the instance with p-value closest to the median of the respective histograms. For these two representatives, we show Kaplan-Meier (KM) survival curves in the top panels of Figure 6(b), respectively. 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.

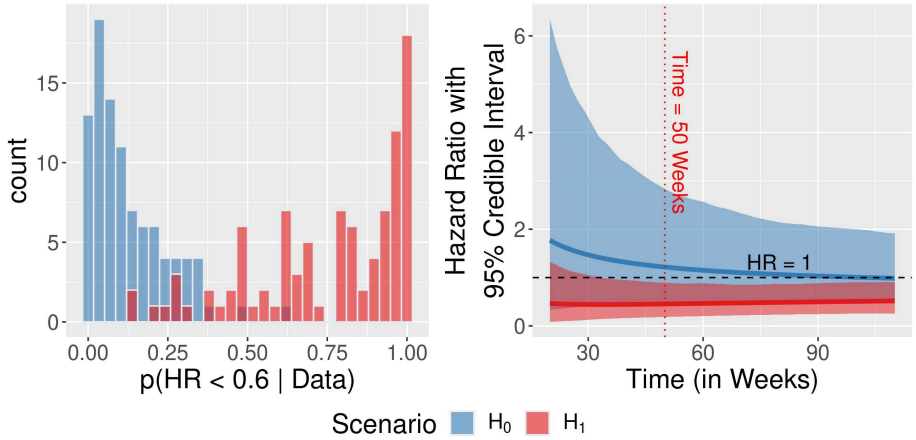
(ii) Matching-based two-step procedure: We use the Matching procedure to create a synthetic control and then follow the same routine of (i) for inference on treatment effects. The results are provided in the bottom panels of Figures 6(a) and 6(b). The distribution of the p-values under the H_1 scenario 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.

(iii) 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 of the CA-PPMx model (see Section S.5 in the supplementary materials for details). Quantile-quantile plots for the two scenarios are shown in Figure 7(a). Near diagonal lines indicate no evidence for a lack of fit. We then evaluate the posterior probability $p_\ell \equiv p(\text{HR} < 0.6 \mid \text{Data})$ (with ℓ indexing the $L = 100$ repeat simulations) at $t = 50$ weeks under the proposed model. The left panel of Figure 7(b) shows histograms of p_ℓ under H_0 (in blue) and under H_1 (in red). As desired, the posterior probabilities are clustered near 0 under H_0, but are peaked near 1 under H_1.

Finally, we identify a representative simulation again by selecting the repeat simulation ℓ with posterior probability p_ℓ closest to the median of the respective histograms



(a) QQ plots for goodness of fit where the $y = x$ line indicates perfect fit.



(b) $p(H_1 \mid \text{Data})$ (left) and hazard ratio with 95% posterior credible interval (right).

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.

under each of the two scenarios. For each of the two scenarios, we plot the posterior estimated hazard ratios (blue and red for simulation under H_0 and H_1, respectively), together with pointwise 95% posterior credible intervals in the right panel of Figure 7(b). Under H_0 (blue), HR is almost equal to 1 with wide credible intervals, whereas under H_1 (red), HR is significantly below 1 with high posterior probability. The median (over the L simulations) posterior probabilities $p_\ell(\text{HR} < 0.6 \mid \text{Data})$ are 0.08 and 0.98 under H_0 and H_1, respectively.

8 Discussion

With a long term goal of setting up a platform for future single-arm early-phase clinical trials in GBM, where new patients only receive experimental therapies, in this article we developed a Bayesian nonparametric approach for creating synthetic controls from RWD. We introduced a Bayesian CAM model that clusters covariates with similar values across different treatment arms.

The flexibility of the CAM model makes it easily generalizable to other problems, e.g., to create two synthetic treatment arms to compare two treatments based on RWD from electronic health records.

Another direction for extensions could build on extracting propensity scores as inference summaries under the CA-PPMx model. This is briefly discussed in Section S.6 of the supplementary materials.

A limitation of the current model is scalability to high-dimensional covariates. In the GBM application, 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 *et al.* (2021) 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.

References

Aldape, K., Brindle, K. M., *et al.* (2019). Challenges to curing primary brain tumours. *Nature Reviews Clinical Oncology*, **16**, 509–520.

Alexander, B. M., Trippa, L., Gaffey, S., *et al.* (2019). Individualized screening trial of innovative Glioblastoma therapy (INSIGhT): A Bayesian adaptive platform trial to develop precision medicines for patients with Glioblastoma. *JCO Precision Oncology*, **3**, 1–13.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika*. To appear.

Au, S. and Beck, J. (2003). Important sampling in high dimensions. *Structural Safety*, **25**, 139–163.

Berry, S. M., Carlin, B. P., Lee, J. J., and Müller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC Press.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Burcu, M., Dreyer, N. A., *et al.* (2020). Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiology and Drug Safety*, **29**, 1228–1235.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, **14**, 1303–1356.

Chandra, N. K., Canale, A., and Dunson, D. B. (2021). Escaping the curse of dimensionality in Bayesian model-based clustering. *arXiv preprint arXiv:2006.02700*.

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.

Chen, W.-C., Wang, C., Li, H., Lu, N., Tiwari, R., Xu, Y., and Yue, L. Q. (2020). Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*, **30**, 508–520.

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

Choi, J., Dekkers, O. M., and le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, **34**, 23–36.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Dahl, D. B. (2006). *Model-based clustering for expression data via a Dirichlet process mixture model*, pages 201–218. Cambridge University Press.

Davi, R., Mahendraratnam, N., Chatterjee, A., *et al.* (2020). Informing single-arm clinical trials with external controls. *Nature Reviews Drug Discovery*, **19**, 821–822.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2021). A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*. To appear.

FDA (2020). Rare diseases at FDA. *https://www.fda.gov/patients/rare-diseases-fda*. Accssed on 7th Dec, 2021.

FDA (2021). *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products*. Guidance for Industry, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products.

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

Fisher, J. P. and Adamson, D. C. (2021). Current FDA-approved therapies for high-grade malignant gliomas. *Biomedicines*, **9**.

Franklin, J. M., Glynn, R. J., Martin, D., and Schneeweiss, S. (2019). Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clinical Pharmacology & Therapeutics*, **105**, 867–877.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **27**, 143–158.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grossman, S. A. and Ellsworth, S. G. (2016). Published glioblastoma clinical trials from 1980 to 2013: Lessons from the past and for the future. *Journal of Clinical Oncology*, **34**, e13522–e13522.

Grossman, S. A., Schreck, K. C., Ballman, K., and Alexander, B. (2017). Point/counterpoint: Randomized versus single-arm phase II clinical trials for patients with newly diagnosed glioblastoma. *Neuro-Oncology*, **19**, 469–474.

Haase, S., Garcia-Fabiani, M. B., *et al.* (2018). Mutant ATRX: Uncovering a new therapeutic target for glioma. *Expert Opinion on Therapeutic Targets*, **22**, 599–613.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, **15**, 609–627.

Hasegawa, T., Claggett, B., *et al.* (2017). The myth of making inferences for an overall treatment efficacy with data from multiple comparative studies via meta-analysis. *Statistics in Biosciences*, **9**, 284–297.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, **67**, 1047–1056.

Jiang, L., Nie, L., and Yuan, Y. (2023). Elastic priors to dynamically borrow information from historical data in clinical trials. *Biometrics*, **79**, 49–60.

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis*, **2**, 719–733.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, **27**, 435–454.

Li, X. and Song, Y. (2020). Target population statistical inference with data integration across multiple sources-an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research*, **12**, 322–333.

Linero, A. R. and Daniels, M. J. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, **33**, 198–213.

Lu, N., Wang, C., Chen, W.-C., Li, H., Song, C., Tiwari, R., Xu, Y., and Yue, L. Q. (2022). Leverage multiple real-world data sources in single-arm medical device clinical studies. *Journal of Biopharmaceutical Statistics*, **32**, 107–123.

Mandel, J. J., Yust-Katz, S., *et al.* (2017). Inability of positive phase II clinical trials of investigational treatments to subsequently predict positive phase III clinical trials in glioblastoma. *Neuro-Oncology*, **20**, 113–122.

Meloun, M. and Militký, J. (2011). The exploratory and confirmatory analysis of univariate data. In *Statistical Data Analysis*, pages 25–71. Woodhead Publishing India.

Miller, J. W. and Harrison, M. T. (2013). A simple example of dirichlet process mixture inconsistency for the number of components. *Advances in Neural Information Processing Systems*, **26**.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, **20**, 260–278.

Müller, P., Chandra, N. K., and Sarkar, A. (2023). Bayesian approaches to include real-world data in clinical studies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **381**, 20220158.

Nam, J. Y. and de Groot, J. F. (2017). Treatment of glioblastoma. *Journal of Oncology Practice*, **13**, 629–638.

Nichol, A., Bailey, M., and Cooper, D. (2010). Challenging issues in randomised controlled trials. *Injury*, **41**, S20–S23.

Ostrom, Q. T., Gittleman, H., *et al.* (2016). CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2009–2013. *Neuro-Oncology*, **18**, v1–v75.

Page, G. L., Quintana, F. A., and Müller, P. (2022). Clustering and prediction with variable dimension covariates. *Journal of Computational and Graphical Statistics*, **31**, 466–476.

Patel, B. and Kim, A. H. (2020). Laser interstitial thermal therapy. *Missouri Medicine*, **117**, 50–55.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 185–207.

Prevost, T. C., Abrams, K. R., and Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine*, **19**, 3359–3376.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, **103**, 1131–1154.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Schmidli, H., Häring, D. A., Thomas, M., Cassidy, A., Weber, S., and Bretz, F. (2020). Beyond randomized clinical trials: Use of external controls. *Clinical Pharmacology & Therapeutics*, **107**, 806–816.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Skare, O., Bølviken, E., and Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, **30**, 719–737.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25**, 1–21.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, **10**, 277–303.

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.

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.

Vansteelandt, S. and Daniel, R. (2014). On regression adjustment for the propensity score. *Statistics in Medicine*, **33**, 4053–4072.

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.

Wang, C., Li, H., Chen, W.-C., *et al.* (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, **29**, 731–748.

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. PMID: 31707908.

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.

Supplementary Materials for

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

Noirrit Kiran Chandra[a] (noirrit.chandra@utdallas.edu)
Abhra Sarkar[b] (abhra.sarkar@utexas.edu)
John F. de Groot[c] (john.degroot@ucsf.edu)
Ying Yuan[d] (yyuan@mdanderson.org)
Peter Müller[b,e] (pmueller@math.utexas.edu)

[a]Department of Mathematical Sciences,
The University of Texas at Dallas, TX, USA

[b]Department of Statistics and Data Sciences,
The University of Texas at Austin, TX, USA

[c]Department of Neurological Surgery,
University of California San Francisco, CA, USA

[d]Department of Biostatistics,
The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[e]Department of Mathematics,
The University of Texas at Austin, TX, USA

Supplementary materials present additional discussion on the motivating dataset, a brief review on the PPMx, detailed discussion on the graphical goodness-of-fit test of our regression model, an alternative interpretation of our model-based inference approach, choices of hyperparameters, detailed posterior simulation scheme, additional simulation studies and associated details, and MCMC convergence diagnostics.

S.1 Historical Data and Potential Future Trial

Figure S.1 shows summaries for the covariates described in Section 2 in the historical database and a potential future single-arm trial. Marginal frequencies for each of the covariates are plotted clearly highlighting the differences between the two populations.

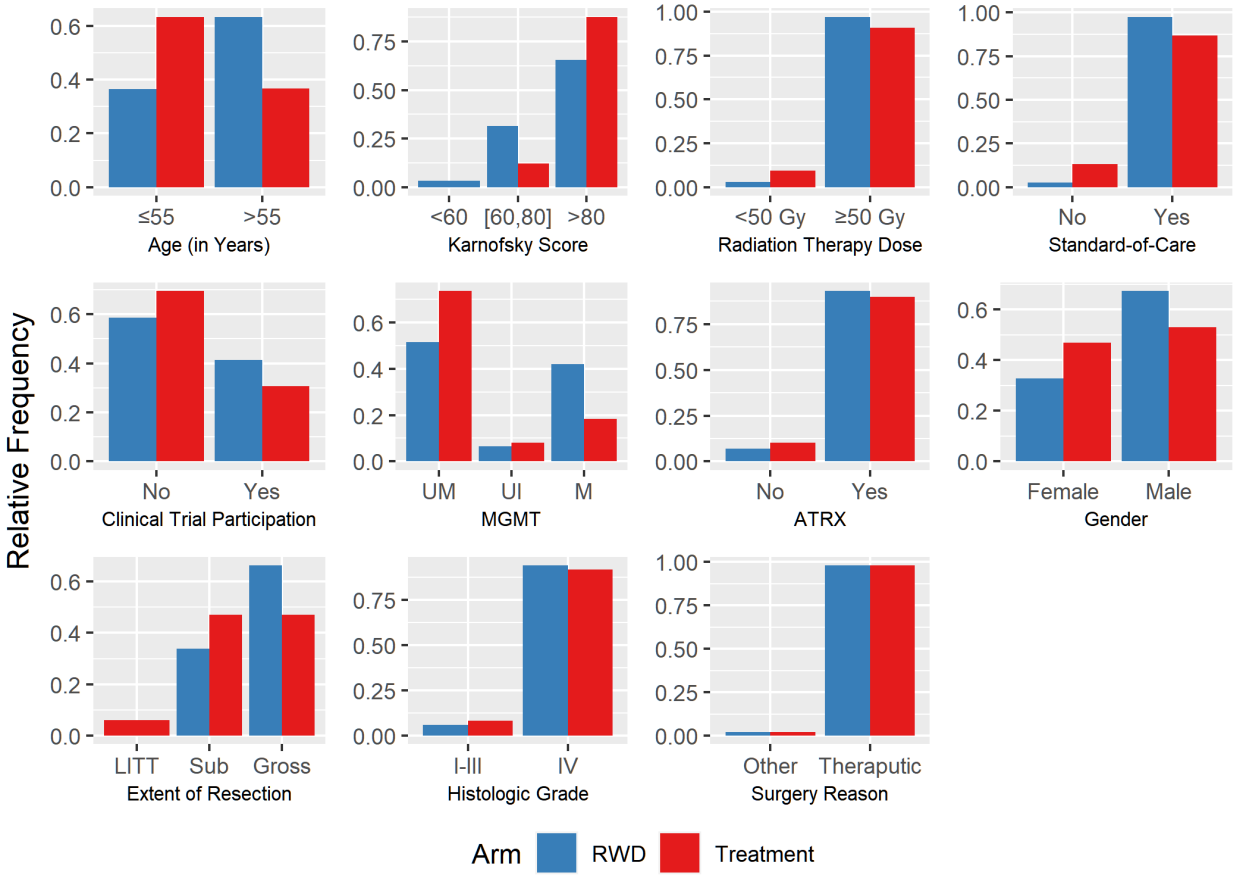


Figure S.1: Relative frequency plots of the covariates in the two treatment arms.

S.2 Product Partition Model with Regression (PPMx)

Let $i = 1, \ldots, n$ be the indices of n data points. For the i^{th} unit (patient, in our case), the data consists of covariates $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ and response variables $\boldsymbol{Y}_i$. Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ be the complete set of covariates and responses respectively. Let $\boldsymbol{\rho}_n = \{S_1, \ldots, S_{k_n}\}$ denote a partition of the n units into k_n subsets, where $1 \leq k_n \leq n$. An equivalent representation of $\boldsymbol{\rho}_n$ introduces cluster membership indicators $c_i = j$ if and only if $i \in S_j$. Let $\boldsymbol{X}_j^{\star}$ be the covariates corresponding to the samples in S_j. In the PPMx, it is believed that data points with more similar covariate values are more

likely to *a priori* be in the same cluster and the corresponding responses are also very similar. The prior consists of two functions - (i) a cohesion function denoted by $c(S_j \mid \alpha) \geq 0$ for $S_j \subset \{1, \ldots, n\}$ associated with a hyper-parameter α discerning the prior belief of co-clustering of the elements of S_j, and (ii) a similarity function denoted by $\boldsymbol{g}(\boldsymbol{X}_j^\star \mid \boldsymbol{\xi})$ and parametrized by $\boldsymbol{\xi}$, formalizing the 'closeness' of the $\boldsymbol{X}_i$'s in the cluster S_j by producing larger values of $\boldsymbol{g}(\boldsymbol{X}_j^\star \mid \boldsymbol{\xi})$ for $\boldsymbol{X}_i$'s that are more similar. Using the similarity and cohesion functions, the PPMx assumes

$$\Pi\left(\boldsymbol{\rho}_n \mid \boldsymbol{X}, \alpha, \boldsymbol{\xi}\right) \propto \prod_{j=1}^{k_n} c(S_j \mid \alpha) \boldsymbol{g}(\boldsymbol{X}_j^\star \mid \boldsymbol{\xi}). \tag{S.1}$$

A default choice for the first factor is $c(S_j \mid \alpha) = \alpha \times (|S_j| - 1)!$, where $\alpha > 0$ and $|\cdot|$ being the cardinality of a set, which is identical to probability function for a random partition under the Chinese restaurant process (Ferguson, 1973). For the second factor, Müller *et al.* (2011) suggested the following default choice for similarity functions

$$\boldsymbol{g}(\boldsymbol{X}_j^\star \mid \boldsymbol{\xi}) = \int \prod_{i \in S_j} q(\boldsymbol{X}_i \mid \boldsymbol{\zeta}_j) G_0(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\zeta}_j. \tag{S.2}$$

With a conjugate sampling model and prior pair of q and G_0, the integral in (S.2) is analytically available, facilitating easy computation. The pair is used to assess the agreement of the data points in S_j rather than any notion of statistical modeling.

The model construction is concluded by specifying a sampling model for the response variable $\boldsymbol{Y}_i$'s. Let $c_i = j$ if $i \in S_j$ denote cluster membership indicators for all $i = 1, \ldots, n$. For a given partition $\boldsymbol{\rho}_n$, we introduce cluster-specific parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_n}\}$ and assume

$$\boldsymbol{Y}_i \mid \boldsymbol{\theta}, c_i = j \overset{\text{ind}}{\sim} h(\boldsymbol{Y}_i \mid \boldsymbol{\theta}_j), \quad \boldsymbol{\theta}_j \mid \boldsymbol{\varphi} \overset{\text{iid}}{\sim} \Pi(\boldsymbol{\theta}_j \mid \boldsymbol{\varphi}), \tag{S.3}$$

where h is a sampling model and $\Pi(\cdot \mid \boldsymbol{\varphi})$ is a prior on $\boldsymbol{\theta}_j$ with possible hyper-parameters $\boldsymbol{\varphi}$.

Recognizing that $\boldsymbol{X}_i$'s may not be random, with slight abuse of notations, under the similarity function (S.2) the PPMx can be equivalently stated as

$$\boldsymbol{X}_i \mid c_i = j, \boldsymbol{\zeta} \overset{\text{iid}}{\sim} q(\boldsymbol{X}_i \mid \boldsymbol{\zeta}_j), \quad \boldsymbol{\zeta}_j \mid \boldsymbol{\xi} \overset{\text{iid}}{\sim} G_0(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}), \quad p(\boldsymbol{\rho}_n) \propto \prod c(S_j \mid \alpha). \tag{S.4}$$

S.3 Missing Data in PPMx

Following the thread of the discussion on handling missing data from Section 3.1 of the main paper, we would like to point out that the model never rules out the possibility of co-clustering a unit with missing entries with fully observed units. For the following argument consider (S.4) with

$$\boldsymbol{X}_i \mid c_i = j, \boldsymbol{\zeta}_j = (\zeta_{j,1}, \ldots, \zeta_{j,p})^{\mathrm{T}} \overset{\text{ind}}{\sim} \textstyle\prod_{\ell=1}^{p} q_\ell(X_{i,\ell} \mid \zeta_{j,\ell}),$$

that is, with $q(\boldsymbol{X}_i \mid \boldsymbol{\zeta}_j)$ factoring over covariates. While implementing inference using a Gibbs sampler, we then update the c_i as follows

$$\Pi(c_i = j \mid \boldsymbol{X}_i, \boldsymbol{\zeta}_{1:K}, \boldsymbol{c}_{-i}) \propto \Pi(c_i = j \mid \boldsymbol{c}_{-i}) \times \textstyle\prod_{\ell=1}^{p} q_\ell(X_{i,\ell} \mid \zeta_{j,\ell}), \tag{S.5}$$

where $\boldsymbol{c}_{-i}$ is the set of c_ℓ's for $\ell = 1, \ldots, n$ excluding c_i.

Now consider the case where we have missing observations in some components of $\boldsymbol{X}_i$ and let $\mathcal{O}_i = \{1 \leq \ell \leq p : X_{i,\ell} \text{ is observed}\}$ be the indices of the observed variables in $\boldsymbol{X}_i$. In this case (S.5) changes to

$$\Pi(c_i = j \mid \boldsymbol{X}_i, \boldsymbol{\zeta}_{1:K}, \boldsymbol{c}_{-i}) \propto \Pi(c_i = j \mid \boldsymbol{c}_{-i}) \times \textstyle\prod_{\ell \in \mathcal{O}_i} q_\ell(X_{i,\ell} \mid \zeta_{j,\ell}).$$

While updating the cluster membership of the units, only the observed variables $X_{i,\ell}$'s in $\boldsymbol{X}_i$ are matched with the corresponding $\zeta_{j,\ell}$ for all $\ell \in \mathcal{O}_i$. A more detailed discussion can be found in Page *et al.* (2022).

S.4 Variations of the Importance Resampling Scheme

S.4.1 Number of Patients to Resample from the RWD

Due to various reasons (see, e.g., Hey and Kimmelman, 2014, for a review), in two-arm designs the allocation of patients in the treatment and control arms are generally considered to be equal, including in particular early-phase GBM trials (Stupp *et al.*, 2014; Nabors *et al.*, 2015; Vanderbeek *et al.*, 2018). As a rule of thumb, we thus recommend the size of the resampled population to be equal to the treatment arm population.

However, if desired any different ratio of sample sizes in treatment and control arm, say $R : 1$, could be used. In that case, even if the the distribution of the covariates in the two arms are same after the importance resampling population adjustment, the AUC of any classifier used in step 5 of Algorithm 1 would be $R/(R+1)$, rather than 0.5.

S.4.2 Averaging over Multiple Resamplings

It may be tempting to average over multiple, say R, instances of the random importance-resampling, to remove one source of variability. But this gives rise to some fundamental problems. For illustrative purpose, we refer to Section 7 of the main manuscript where we discuss the application in GBM. There we use the importance resampling strategy to generate an equivalent subpopulation of the treatment arm and then use the Cox proportional hazard model to test for treatment effects. In Figure 6(a), we plot the histogram of p-values under the null scenario which resembles the Unif$(0, 1)$ distribution. Now for R resamplings we would

have multiple p-values corresponding to each of the R resampled populations. Subsequently we need a statistic to summarize the p-values, let us denote it by T. Letting $p_1, \ldots, p_R$ be the p-values thus obtained, the distribution of $T(p_1, \ldots, p_R)$ will not be $U(0,1)$ anymore under the null. We therefore recommend against it. As importance resampling schemes are asymptotically unbiased (Skare *et al.*, 2003), under reasonably large sample sizes, a single resampled population should be adequate.

S.5 Goodness-of-Fit Test for Continuous Responses

We use the approach of Johnson (2007) to suggest a graphical goodness-of-fit tool to validate the mixture of lognormals model for the CA-PPMx. The procedure is valid as long as long as h in (7) is a univariate continuous density, i.e., as long as the response variables are univariate and continuous. For the moment, we suppress the additional s subindex on $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$. Let $m(\boldsymbol{Y} \mid \boldsymbol{X})$ be the marginal distribution after integrating out all model parameters

$$m(\boldsymbol{Y} \mid \boldsymbol{X}) = \sum_{\boldsymbol{c}} \int \left\{ \prod_{i=1}^{n} h(Y_i \mid \boldsymbol{\theta}_{c_i}) \right\} \mathrm{d}p(\boldsymbol{\theta}, \boldsymbol{c}_{1:n} \mid \boldsymbol{X}).$$

We implement a test of fit based on the following result. Assuming that $m(\boldsymbol{Y} \mid \boldsymbol{X})$ is the true marginal distribution of $\boldsymbol{Y}$, we have:

Proposition 1. *Let $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{c}_{1:n})$ be a sample from their posterior, $H(y \mid \boldsymbol{\theta}) = \int_{-\infty}^{y} h(z \mid \boldsymbol{\theta})\mathrm{d}z$ be the CDF, and $U_i = H(Y_i \mid \boldsymbol{\theta}_{c_i})$, $i = 1, \ldots, n$. Then, $U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$.*

Proof. Let $\boldsymbol{u}_{1:n} = \{u_1, \ldots, u_n\}$ and define $A(\boldsymbol{u}_{1:n}; \boldsymbol{\omega}) = \cap_{i=1}^{n} \{y : H(y \mid \boldsymbol{\theta}_{c_i}) \leq u_i\}$. Then,

$$\Pr(U_i \leq u_i \text{ for all } i = 1, \ldots, n) = \int \int_{A(\boldsymbol{u}_{1:n}; \boldsymbol{\omega})} \mathrm{d}\Pi(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{Y}) m(\boldsymbol{Y} \mid \boldsymbol{X}) \mathrm{d}\boldsymbol{Y}.$$

Note that $\Pi(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{Y}) = \{\prod_{i=1}^{n} h(Y_i \mid \boldsymbol{\theta}_{c_i})\} \Pi(\boldsymbol{\omega} \mid \boldsymbol{X}) / m(\boldsymbol{Y} \mid \boldsymbol{X})$. Substituting this in the above equation, we get

$$\Pr(U_i \leq u_i \text{ for all } i = 1, \ldots, n) = \int \left\{ \int_{A(\boldsymbol{u}_{1:n}; \boldsymbol{\omega})} \prod_{i=1}^{n} h(Y_i \mid \boldsymbol{\theta}_{c_i}) \mathrm{d}Y \right\} \mathrm{d}\Pi(\boldsymbol{\omega} \mid \boldsymbol{X}).$$

Now, the term inside the parenthesis integrates to $\prod_{i=1}^{n} u_i$ which is independent from $\Pi(\boldsymbol{\omega} \mid \boldsymbol{X})$. Hence the proof. □

To understand the implications, consider the distribution $(\boldsymbol{Y}, \boldsymbol{\omega} \mid \boldsymbol{X})$ for a hypothetical data set $(\boldsymbol{X}, \boldsymbol{Y})$. First sample $\widetilde{\boldsymbol{\omega}} = (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{c}}_{1:n})$ from $p(\boldsymbol{\omega} \mid \boldsymbol{X}) = p(\boldsymbol{c} \mid \boldsymbol{X})\ p(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{X})$ and then $(\boldsymbol{Y} \mid \widetilde{\boldsymbol{\omega}}, \boldsymbol{X})$ from the sampling model (7). Letting $\widetilde{U}_i = H(Y_i \mid \widetilde{\boldsymbol{\theta}}_{\widetilde{c}_i})$, we then have $\widetilde{U}_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Assuming that the observed data $\boldsymbol{Y}$ do in fact arise from the assumed marginal model $m(\boldsymbol{Y} \mid \boldsymbol{X})$, Proposition 1 sets up sampling from the alternative factorization $p(\boldsymbol{Y}, \boldsymbol{\omega} \mid \boldsymbol{X}) = m(\boldsymbol{Y} \mid \boldsymbol{X}) \cdot p(\boldsymbol{\omega} \mid \boldsymbol{Y}, \boldsymbol{X})$. It follows that $\widetilde{\boldsymbol{U}}_{1:n}$ and $\boldsymbol{U}_{1:n}$ are indistinguishable in distribution. The latter, $\boldsymbol{U}_{1:n}$, can be readily obtained from the posterior samples of

$\boldsymbol{\omega}$. Letting $\boldsymbol{U}_{1:n}^{(m)}$ denote the evaluation under the m^{th} posterior MCMC sample $\boldsymbol{\omega}^{(m)}$, a goodness-of-fit test can then be carried out to validate the uniform distribution.

Note that the $\boldsymbol{U}_{1:n}^{(m)}$'s vary across different posterior samples $\boldsymbol{\omega}^{(m)}$ while also having hierarchical dependence since all of them are sampled conditionally on the same $\boldsymbol{Y}$ (and $\boldsymbol{X}$). Although in principle formal prior-predictive-posterior based tests be carried out (Johnson, 2007; Cao *et al.*, 2010), it can be numerically infeasible for complex models like ours. As a practical alternative, goodness-of-fit can be assessed by inspecting the quantile-quantile plots of $\boldsymbol{U}_{1:n}^{(m)}$. Such visual tools can be effective for detecting departures from model assumptions (Meloun and Militký, 2011, Chapter 2). We use it to assess the model fit in Section 7.

To assess the goodness-of-fit in the GBM application, where the outcomes are right-censored survival data, we extend the result in the following corollary.

Corollary 1. *Suppose we have right-censored survival outcomes (Y_i, ν_i) with covariate $\boldsymbol{X}_i$ where $\nu_i = 1$ if Y_i is an observed failure time, for $i = 1, \ldots, n$. Following the notations of Theorem 1, define $U_i = H(Y_i \mid \boldsymbol{\theta}_{c_i})$ if $\nu_i = 1$, else if $\nu_i = 0$ define $U_i = H(Y_i \mid \boldsymbol{\theta}_{c_i}) + \gamma_i\{1 - H(Y_i \mid \boldsymbol{\theta}_{c_i})\}$, where $\gamma_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ independent from Y_i. If the observed failure times are independent of the censoring times, then $U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$.*

Proof of Corollary 1. Let $\widetilde{Y}_i$ be the true failure time of the i^{th} individual, that is $\widetilde{Y}_i \geq Y_i$ with equality if and only if $\nu_i = 1$. Letting $\widetilde{U}_i = H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i})$, Theorem 1 implies $\widetilde{\boldsymbol{U}}_{1:n} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Note that

$$H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i}) = \nu_i H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i}) + (1-\nu_i)\left[H(Y_i \mid \boldsymbol{\theta}_{c_i}) + \{H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i}) - H(Y_i \mid \boldsymbol{\theta}_{c_i})\}\right].$$

Since $H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i}) \sim \text{Unif}(0,1)$ and is independent of Y_i, $H(Y_i \mid \boldsymbol{\theta}_{c_i}) + \{H(\widetilde{Y}_i \mid \boldsymbol{\theta}_{c_i}) - H(Y_i \mid \boldsymbol{\theta}_{c_i})\} \mid Y_i, \boldsymbol{\theta}_{c_i} \sim \text{Unif}\{H(Y_i \mid \boldsymbol{\theta}_{c_i}), 1\}$ which follows the same distribution as $\gamma_i\{1 - H(Y_i \mid \boldsymbol{\theta}_{c_i})\}$. Hence the proof. □

S.5.1 Illustrating Example for the Graphical Goodness-of-Fit Test

We illustrate the Bayesian goodness-of fit test in a linear regression problem. We simulate data $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$ $(= 1,000)$ from the following mixture distribution

$$Y_i \mid \boldsymbol{X}_i \overset{\text{ind}}{\sim} \pi_0 \text{N}(\alpha_0 + \boldsymbol{\beta}_0^{\text{T}}\boldsymbol{X}_i, \sigma_0^2) + (1-\pi_0)\text{Exp}(\alpha_0 + \boldsymbol{\beta}_0^{\text{T}}\boldsymbol{X}_i), \tag{S.6}$$

where $\boldsymbol{X}_i$'s are p $(= 5)$-variate continuous covariates and $\text{Exp}(a)$ denotes an exponential distribution with mean a. However, we fit the following misspecified Bayesian linear regression model on the data using the `MCMCpack R` package

$$\begin{aligned} &\textbf{likelihood: } Y_i \mid \boldsymbol{X}_i \overset{\text{ind}}{\sim} \text{N}(\alpha + \boldsymbol{\beta}^{\text{T}}\boldsymbol{X}_i, \sigma^2); \\ &\textbf{prior: } (\alpha, \boldsymbol{\beta}) \sim \text{N}_{p+1}(\boldsymbol{0}, 10 \times \boldsymbol{I}_{p+1}),\ \sigma^{-2} \sim \text{Ga}(0.1, 0.1). \end{aligned} \tag{S.7}$$

For varying values of π_0, we show quantile-quantile plots in Figure S.2 where we see deviation from the diagonal $y = x$ straight-line aggravates as $\pi_0 \to 0$, i.e., with increasing model misspecification.

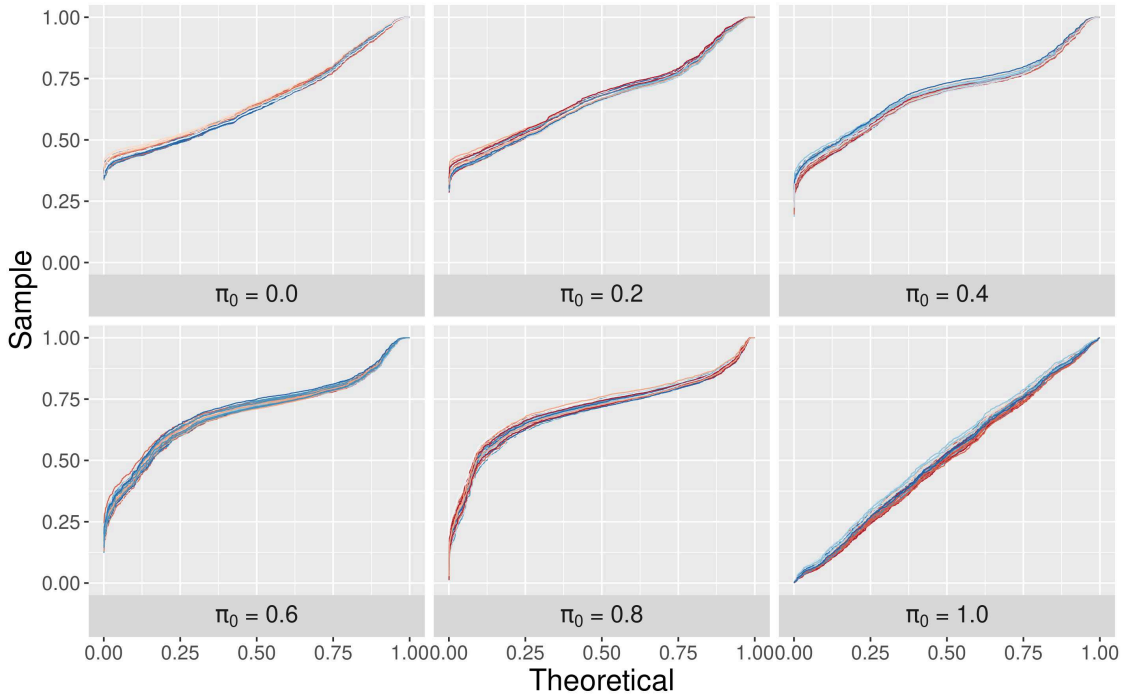


Figure S.2: Quantile-quantile plots for increasing model misspecification: Data are generated from model (S.6) for different values of π_0 and the Bayesian linear regression model in Eqn (S.7) is fitted where $\pi_0 = 0.0$ and 1.0 denote the extreme misspecified model and the true model, respectively. Deviation from the diagonal $y = x$ straight-line aggravates with increasing model misspecification.

S.6 Alternate Interpretation of the CA-PPMx

In Section 4.2, we introduced a model-based approach for inference on treatment effects in the CA-PPMx model. An alternative interpretation of the approach arises from observing the following connection with methods based on PS stratification (Wang *et al.*, 2019; Chen *et al.*, 2020; Lu *et al.*, 2022). The CAM model can be interpreted as a stochastic PS stratification. To see this, first re-index all patients and patient specific variables across $s = 1, 2$ as $i = 1, \ldots, N = n_1 + n_2$ and define $Z_i \in \{1, 2\}$ if patient i was originally in data set $s = 1$ or 2, respectively. Assuming equal sample sizes $n_1 = n_2$, we have $p(Z_i = 1 \mid c_i = j)/p(Z_i = 2 \mid c_i = j) = \pi_{1,j}/\pi_{2,j}$. That is, the terms in the CAM model correspond to different PS ratios for the selection of a patient into $s = 1$ versus $s = 2$. Grouping patients in clusters C_j is then interpreted as stratification by PS, with clusters C_j defining the strata. Within each stratum we report treatment effect $\delta_j = \delta\{h(Y \mid \boldsymbol{\theta}_{1,j}), h(Y \mid \boldsymbol{\theta}_{2,j})\}$. Compare the discussion in Section 4.2.

Whereas fixed consolidated unidimensional PSs may be inadequate in matching multivariate covariates (Stuart, 2010; King and Nielsen, 2019) and hence sensitive to the specification of the PS model (Zhao, 2004), inference under the proposed CAM model overcomes limitations by naturally including uncertainty in the stratification.

S.7 CA-PPMx Specifications and Hyperparameters

Recall the setup from Section 3.1 and the notations from Eqn (4). For categorical covariate $X_{s,\ell}$ with categories $1,\ldots,m_\ell$, we choose $q_\ell(X_{s,\ell} \mid \boldsymbol{\zeta}_\ell) = \text{Mult}(1;\zeta_{\ell,1},\ldots,\zeta_{\ell,m_\ell})$ and $g_{0,\ell}(\zeta_{\ell,1},\ldots,\zeta_{\ell,m_\ell}) = \text{Dir}(1,\ldots,1)$ to choose a uniform distribution over the simplex. For continuous $X_{s,\ell}$, we choose $q_\ell(X_{s,\ell} \mid \boldsymbol{\zeta}_\ell) = \text{N}(X_{s,\ell};\mu_{X,\ell},\sigma^2_{X,\ell})$ with $\boldsymbol{\zeta}_\ell = (\mu_{X,\ell},\sigma^2_{X,\ell})$ and $g_{0,\ell}(\mu_{X,\ell},\sigma^2_{X,\ell}) = \text{NIG}(\mu_{X,\ell},\sigma^2_{X,\ell};0,1,\alpha_X,1)$, i.e., $\mu_{X,\ell} \mid \sigma^2_{X,\ell} \sim \text{N}(0,\sigma^2_{X,\ell})$, $\sigma^{-2}_{X,\ell} \sim \text{Ga}(a_X,1)$. Following standard practice, we center $\mu_{X,\ell}$ around zero. Based on previous experience on Gaussian mixture models, we set $a_X = \#\text{continuous covariates} + 30$, as a small prior variance on $\sigma^2_{X,\ell}$'s favors a larger number of occupied clusters in the mixture model a posteriori, allowing for a more flexible fit. Recall that we have assumed $\log\alpha_s \sim \text{N}(\mu_\alpha,\sigma^2_\alpha)$ for $s=1,2$ on the concentration parameters in models (1) and (3). To specify weakly informative priors, we set the hyperparameters μ_α and σ^2_α such that $\mathbb{E}(\alpha_s)=1$ and $\text{var}(\alpha_s)=10$ a priori for $s=1,2$.

Regarding the parameters of the sampling model for survival outcomes in Eqn (8), we set $\kappa_0=1$ and $a_0=10$ to ensure a thin-tailed base-measure. In our experience, with too heavy tailed prior distributions, small sample performance can easily get dominated by the prior. Regarding the hyperprior on the mean parameter μ_0, we choose m_μ using an empirical Bayes type approach. Letting $\widetilde{n}$ be the number of observed failures combining the RWD and the current trial, we set $m_\mu = \frac{1}{\widetilde{n}}\sum_s\sum_{i:\nu_{s,i}=1} Y_{s,i}$, i.e., the grand mean of the log-observed failure times across all arms. We further set $s^2_\mu = 1$. Regarding the hyperprior on the scale parameter b_0, we choose m_b and s^2_b such that $\mathbb{E}(b_0)=5$ and $\text{var}(b_0)=20$ a priori to set a weakly informative hyperprior.

Regarding the real-valued continuous responses in the simulation studies in Section 6, we use the model in Eqn (8) on the actual response variables with $\nu_{s,i}=1$ for all i and s.

S.8 Posterior Computation

For computational convenience in the practical implementation, we consider the *degree k weak limit approximation* (Ishwaran and Zarepour, 2002a,b) of the GEM(α_2) distribution in (1), i.e., we use a Dir($\alpha_2/k,\ldots,\alpha_2/k$) distribution, with fixed but large enough k. We set $k=15$ for all our simulation experiments and applications.

We develop a Gibbs sampler to avoid computational issues with a Gaussian mixture models on the log transformed survival outcomes with censoring. Without loss of generality we assume $Y_{s,i}$'s (log transformed outcomes) are supported on the entire real line and describe our algorithm for a mixture of Gaussian distributions. Let $\nu_{s,i}$'s be the censoring indicators such that $\nu_{s,i}=1$ implies $Y_{s,i}$ is an observed failure time; else if it is censored in the interval $(Y_{s,i,l},Y_{s,i,u})$ then $\nu_{s,i}=0$. For left and right censoring, we take $Y_{s,i,u}=\infty$ and $Y_{s,i,l}=-\infty$, respectively. Let $\widetilde{Y}_{s,i}$ be the true failure times, that is $\widetilde{Y}_{s,i}=Y_{s,i}$ if and only if $\nu_{s,i}=1$. Offline, before starting MCMC simulation, we initialize $\widetilde{Y}_{s,i}$ at some admissible value for $\nu_{s,i}=0$

and cluster membership indicator variables $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. For the CAM model on covariates, we consider a conjugate pair q_ℓ and $g_{0,\ell}$ for $\ell = 1, \ldots, p$. This allows us to analytically marginalize with respect to the atoms $\boldsymbol{\zeta}_j$'s. This strategy results in substantially improved mixing of the Markov chain.

The sampler iterates through the following steps. In Step 1, we impute $\widetilde{Y}_{s,i}$'s for the censored observations; in Step 2, we update the cluster membership indicators $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$; in Step 3, we update hyper-parameters related to the response model that allows sharing of information via a hierarchical model; in Step 4, we update the parameters required to implement the strategies outlined in Sections 3.2 and 4.2; finally in Step 5 we update the Dirichlet hyperparameters for the two mixture models.

Step 1 We define the set $S_{s,j,-i} = \{i : c_{s,i} = j\} \setminus \{i\}$, $n_{s,j,-i} = |S_{s,j,-i}|$, $\kappa_{s,j,-i} = \kappa_0 + n_{s,j,-i}$, $\overline{Y}_{s,j,-i} = \sum_{r \in S_{s,j,-i}} \widetilde{Y}_{s,r}/n_{s,j,-i}$, $\mu_{s,j,-i} = (\kappa_0\mu_0 + n_{s,j,-i}\overline{Y}_{s,j,-i})/\kappa_{s,j,-i}$, $a_{s,j,-i} = a_0 + n_{s,j,-i}/2$, $b_{s,j,-i} = b_0 + \sum_{r \in S_{s,j,-i}} (\widetilde{Y}_{s,r} - \overline{Y}_{s,j,-i})^2/2 + n_{s,j,-i}\kappa_0(\overline{Y}_{s,j,-i} - \mu_0)^2/\kappa_{s,j,-i}$. Then for all $i = 1, \ldots, n_s$ and $s = 1, 2$, generate

$$\widetilde{Y}_{s,i} \sim \begin{cases} Y_{s,i} \text{ with probability 1 if } \nu_{s,i} = 1; \\ t_{2a_{s,j,-i}} \left\{ \mu_{s,j,-i}, \frac{b_{s,j,-i}(\kappa_{s,j,-i}+1)}{a_{s,j,-i}\kappa_{s,j,-i}} \mid (Y_{s,i,l}, Y_{s,i,u}) \right\}, \end{cases}$$

where $t_{df}\{\mu, \sigma^2 \mid (a, b)\}$ is a central Student's t-distribution, with degrees of freedom df, median μ and scale parameter σ, truncated to the set (a, b).

Step 2 Letting $f_t\{\cdot \mid df, \mu, \sigma^2\}$ and $F_t\{\cdot \mid df, \mu, \sigma^2\}$ denote the pdf and cdf of a central Student's t-distribution with degrees of freedom df, median μ and scale parameter σ, respectively, we define

$$\psi_{Y;s,j}(i) = \begin{cases} f_t \left\{ Y_{s,i} \mid 2a_{s,j,-i}, \mu_{s,j,-i}, \frac{b_{s,j,-i}(\kappa_{s,j,-i}+1)}{a_{s,j,-i}\kappa_{s,j,-i}} \right\} \text{ if } \nu_{s,i} = 1; \\ F_t \left\{ Y_{s,i,u} \mid 2a_{s,j,-i}, \mu_{s,j,-i}, \frac{b_{s,j,-i}(\kappa_{s,j,-i}+1)}{a_{s,j,-i}\kappa_{s,j,-i}} \right\} \\ \qquad - F_t \left\{ Y_{s,i,l} \mid 2a_{s,j,-i}, \mu_{s,j,-i}, \frac{b_{s,j,-i}(\kappa_{s,j,-i}+1)}{a_{s,j,-i}\kappa_{s,j,-i}} \right\} \text{ otherwise.} \end{cases}$$

Recall from Section 3.1 (see page 11) that $\mathcal{O}_{s,i}$ is the set of indices of the covariates observed for $\boldsymbol{X}_{s,i}$, and define the sets $\mathcal{C}_{j,\ell} = \cup_{s=1}^2 \{i : i \in S_j, \ell \in \mathcal{O}_{s,i}\}$ and $\boldsymbol{X}^{*o}_{j,\ell} = \cup_{s=1}^2 \{X_{s,i,j} : i \in \mathcal{C}_{j,\ell}\}$. Define the functions $g_\ell(\boldsymbol{X}^{*o}_{j,\ell} \mid \boldsymbol{\xi}_\ell) = \int \prod_{i \in \mathcal{C}_{j,\ell}} q_\ell(X_{s,i,\ell} \mid \boldsymbol{\zeta}_{j,\ell}) g_{0,\ell}(\boldsymbol{\zeta}_{j,\ell} \mid \boldsymbol{\xi}_\ell) \mathrm{d}\boldsymbol{\zeta}_{j,\ell}$ and $\psi_{X;s,j}(i) = \prod_{\ell \in \mathcal{O}_{s,i}} \frac{g_\ell(\boldsymbol{X}^{*o}_{j,\ell} \mid \boldsymbol{\xi}_\ell)}{g_\ell[\boldsymbol{X}^{*o}_{j,\ell} \setminus \{X_{s,j,\ell}\} \mid \boldsymbol{\xi}_\ell]}$. Then, $\boldsymbol{c}_1$ can be updated as

$$\Pi(c_{1,i} = j \mid -) \propto (n_{1,j,-i} + \alpha_1/k(n_2)) \times \psi_{Y;1,j}(i) \times \psi_{X;1,j}(i) \text{ for } j = 1, \ldots, k(n_2).$$

Similarly $\boldsymbol{c}_2$ can be updated as

$$\Pi(c_{2,i} = j \mid -) = 1 \text{ if } n_{1,j} > 0 \text{ and } n_{2,j,-i} = 0;$$
$$\text{else } \Pi(c_{2,i} = j \mid -) \propto (n_{2,j,-i} + \alpha_2/k) \times \psi_{Y;2,j}(i) \times \psi_{X;2,j}(i) \text{ for } j = 1, \ldots, k.$$

Step 3 Define $\widetilde{b} = \log b_0$ and let $\Pi(\mu_0, \widetilde{b} \mid \widetilde{\boldsymbol{Y}}_{1,1:n_1}, \widetilde{\boldsymbol{Y}}_{2,1:n_2})$ be the joint posterior density of μ_0 and $\widetilde{b}$ given $\widetilde{Y}_{s,i}$'s, $k_{n,1}$ and $k_{n,2}$ be the number of non-empty clusters in the two cohorts respectively. Then,

$$\log \Pi(\mu_0, \widetilde{b} \mid \widetilde{\boldsymbol{Y}}_{1,1:n_1}, \widetilde{\boldsymbol{Y}}_{2,1:n_2}) = K - \frac{(\mu_0 - m_\mu)^2}{2s_\mu^2} - \frac{(\widetilde{b} - m_b)^2}{2s_b^2} + (k_{n,1} + k_{n,2})a_0\widetilde{b}$$
$$- \sum_{j=1}^{k(n_2)} \sum_{s=1}^{2} \left(a_0 + \frac{n_{s,j}}{2}\right) \log \left[e^{\widetilde{b}} + \frac{1}{2} \left\{ \mu_0^2\kappa_0 + \sum_{i \in S_{s,j}} \widetilde{Y}_{s,i}^2 - \frac{(\kappa_0\mu_0 + n_{s,j}\overline{Y}_{s,j})^2}{\kappa_0 + n_{s,j}} \right\} \right],$$

where K is a constant and $\overline{Y}_{s,j} = \sum_{i \in S_{s,j}} \widetilde{Y}_{s,i}$. We sample μ_0 and $\widetilde{b}$ using a Hamiltonian Monte Carlo (HMC) algorithm (Duane *et al.*, 1987).

Step 4 For $j = 1, \ldots, k$, we define the set $S_{s,j} = \{i : c_{s,i} = j\}$, $\kappa_{s,j} = \kappa_0 + n_{s,j}$, $\mu_{s,j} = (\kappa_0\mu_0 + n_{s,j}\overline{Y}_{s,j})/\kappa_{s,j}$, $a_{s,j} = a_0 + n_{s,j}/2$, $b_{s,j} = b_0 + \sum_{r \in S_{s,j}} (\widetilde{Y}_{s,r} - \overline{Y}_{s,j})^2/2 + n_{s,j}\kappa_0(\overline{Y}_{s,j} - \mu_0)^2/\kappa_{s,j}$. Then,

$$\mu_{s,j} \sim t_{2a_{s,j}} \left\{ \mu_{s,j}, \frac{b_{s,j}(\kappa_{s,j} + 1)}{a_{s,j}\kappa_{s,j}} \right\}, \qquad \sigma_{s,j}^{-2} \sim \mathrm{Ga}(a_{s,j}, b_{s,j}), \tag{S.8}$$

$$\boldsymbol{\pi}_1 \sim \mathrm{Dir}\left(n_{1,1} + \frac{\alpha_1}{k(n_2)}, \ldots, n_{1,k(n_2)} + \frac{\alpha_1}{k(n_2)}\right), \quad \boldsymbol{\pi}_2 \sim \mathrm{Dir}\left(n_{2,1} + \frac{\alpha_2}{k}, \ldots, n_{2,k} + \frac{\alpha_2}{k}\right).$$

For $s = 1$, we only sample for $j = 1, \ldots, k(n_2)$ in (S.8). Note that the dimension of $\boldsymbol{\pi}_1$ can vary across MCMC samples.

Step 5 With lognormal priors on the Dirichlet mixture hyperparameters α_1 and α_2, $\log \alpha_s \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$, $s = 1, 2$, the log-posterior pdfs are given by

$$\log \Pi(\alpha_1 \mid -) = K_1 + \log \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + n_1)} + \sum_{j : n_{1,j} > 0} \log \frac{\Gamma(\alpha_1/k(n_2) + n_1)}{\Gamma(\alpha_1)} - \log \alpha_1 - \frac{(\log \alpha_1 - \mu_\alpha)^2}{2\sigma_\alpha^2},$$

$$\log \Pi(\alpha_2 \mid -) = K_2 + \log \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_2 + n_2)} + \sum_{j : n_{2,j} > 0} \log \frac{\Gamma(\alpha_2/k + n_2)}{\Gamma(\alpha_2)} - \log \alpha_2 - \frac{(\log \alpha_2 - \mu_\alpha)^2}{2\sigma_\alpha^2}.$$

As the respective pdfs are differentiable with respect to α_1 and α_2, we sample the parameters using HMC.

Remark 1. *Note that in Step 2, $\mathcal{C}_{j,\ell}$ is the set of data points in S_j with observed covariate ℓ, $\boldsymbol{X}_{j,\ell}^{*o}$ is the collection of the observed values of the covariate ℓ in S_j and $g_\ell(\boldsymbol{X}_{j,\ell}^{*o} \mid \boldsymbol{\xi}_\ell)$ is the joint marginal density. A conjugate pair q_ℓ and $g_{0,\ell}$ ensures the analytical availability of g_ℓ and $\psi_{X;s,j}(i)$ becomes the conditional distribution of $\boldsymbol{X}_{s,i}$ given $\boldsymbol{X}_{j,\ell}^{*o}$. For continuous real-valued $X_{s,j,\ell}$, we may take $q_\ell(\cdot \mid \boldsymbol{\zeta}_j)$ to be the univariate Gaussian pdf where $\boldsymbol{\zeta}_j$ is the set of associated mean and variance parameters, and $g_{0,\ell}(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_\ell)$ to be a normal-inverse-gamma density (compare Section S.7). In this case, the ratio $\frac{g_\ell(\boldsymbol{X}_{j,\ell}^{*o} \mid \boldsymbol{\xi}_\ell)}{g_\ell[\boldsymbol{X}_{j,\ell}^{*o} \setminus \{X_{s,j,\ell}\} \mid \boldsymbol{\xi}_\ell]}$ reduces to a central t-distribution density; for categorical $X_{s,j,\ell}$, a convenient choice can be the multinomial-Dirichlet pair which again yields an analytical expression of the ratio.*

In the GBM application and simulation studies in Section 6, we have considered conjugate normal-inverse-gamma and multinomial-Dirichlet conjugate pairs for continuous real-valued

covariates and categorical covariates, respectively. For all simulation studies and GBM application, we consider 6,000 MCMC iterations, discarded the first 1,000 as the burn-in samples, and saved every 5^{th} MCMC sample to reduce autocorrelation.

Finally we note that the complete conditional for $\pi_{1,j}$ in step 4 could be used to implement Rao-Blackwellization (Robert and Roberts, 2021) in the evaluation of the weights w_i in (6) by replacing $\pi_{1,j}$ with the conditional posterior means.

S.9 Additional Details on Simulation Studies

S.9.1 Procedure to Test for Treatment Effects in Section 6

Recall that in Section 6 we test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ in each simulation setup. To compute the power, we first estimate the treatment effect, say $\widehat{\delta}$ in each setup. Estimated treatment effects under CA-PPMx are evaluated using the posterior mean of Eqn (9). To evaluate type-II error rates we use the empirical distribution of $\widehat{\delta}$ under simulation truth $\delta = 0$ for each of the seven methods under consideration across the 500 repeat simulations to obtain their distributions under H_0. We evaluate the empirical 2.5% and 97.5% quantiles, say $\widehat{\delta}_L$ and $\widehat{\delta}_U$ and define the test function $\Phi(\widehat{\delta}) = \mathbb{1}_{\widehat{\delta} \notin [\widehat{\delta}_L, \widehat{\delta}_U]}$ controlling the type-I error at 5% level of significance.

S.9.2 Details on Simulation Truths

CAM scenario: We set $\mu_{1,1} = \mu_{1,1} = 2$ and $\mu_{1,j} = 0$ for all $j > 2$, and $\mu_{2,5} = \mu_{2,6} = 2$ and $\mu_{2,j} = 0$ for all $j \notin \{5, 6\}$, $\sigma_j^2 = 0.05$ for all $j = 1, 2, 3$. Regarding the mixture weights, we set $\pi_{1,1} = \pi_{1,2} = 0.5$ and $\pi_{2,1} = \pi_{2,2} = 1/6$ and $\pi_{2,3} = 2/3$. Regarding the categorical covariates we set $\varrho_1 = 0.85$, $\varrho_2 = 0.65$.

MIX scenario: We take $k = 4$. Recall that $\boldsymbol{\mu}_{1,j} = \boldsymbol{\mu}_{2,j}$ for all $j < k$, say $\boldsymbol{\mu}_j = (\mu_{j,1}, \ldots, \mu_{j,p})^{\mathrm{T}}$. For each $j < k$, we take $\mu_{j,2j+1} = \mu_{j,2j+2} = 2$ and $\mu_{j,\ell} = 0$ for all $\ell \notin \{2j + 1, 2j + 2\}$. Finally for $\boldsymbol{\mu}_{s,k} = (\mu_{s,k,1}, \ldots, \mu_{s,k,p})^{\mathrm{T}}$ with $s = 1, 2$, we set $\mu_{1,k,7} = \mu_{1,k,8} = 2$, $\mu_{1,k,9} = 1$ and $\mu_{1,k,\ell} = 0$ for all $\ell \notin \{7, 8, 9\}$; and $\mu_{2,k,2k+1} = \mu_{2,k,2k+2} = 2$ and $\mu_{2,k,\ell} = 0$ for all $\ell \notin \{2k + 1, 2k + 2\}$. In each repeat simulation we generate $w_{1,1}, \ldots, w_{1,k} = \mathrm{SRSWR}_k(1, \ldots, 4)$ where $\mathrm{SRSWR}_r(\mathcal{S})$ denotes the simple random sampling scheme with replacement of size r from the set $\mathcal{S}$. Then we set $\pi_{1,j} = w_{1,j} / \sum_{r=1}^{k} w_{1,r}$ for all $j = 1, \ldots, k$. we set $\pi_{2,j} = 1/k$ for all $j = 1, \ldots, k$.

Interaction scenario: Recall the covariates in the GBM dataset from Table 2 in the main manuscript. We consider pairwise interactions between (Gender, Age) and (RT Dose, Age). Following that, we have one-hot-encoded the covariates with more than two categories

(e.g., KPS) so that we are left with all binary covariates (including the interactions). Let $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ be the covariates corresponding to patient record i with p being the number of covariates.

For each repeat simulation, we then generate $\boldsymbol{b} = (b_1, \ldots, b_p)^{\mathrm{T}} = \mathrm{SRSWR}_p(-1, 0.75)$. We then assign the patient record i to the treatment arm with probability $\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{b}+0.8}{1+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{b}+0.8}$.

Oracle scenario: We follow the exact same strategy as described in the Interaction scenario but without pairwise interactions.

Outcome model: For $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, we take $f(\boldsymbol{x}) = \beta_1 \mathbb{1}_{(x_1 \geq 1.25, x_2 \geq 1.25)} - \beta_2 \mathbb{1}_{(x_3 \geq 1.25, x_4 \geq 1.25)} + \beta_3 \mathbb{1}_{(x_5 \geq 1.25, x_6 \geq 1.25)} + \beta_4 \mathbb{1}_{(x_{p-1} \geq 1, x_p \geq 1)}$. In each repeat simulation we let $\beta_1, \beta_2 \overset{\text{iid}}{\sim} \mathrm{Unif}(40, 60)$, $\beta_3 \sim \mathrm{Unif}(225, 275)$ and $\beta_4 \sim \mathrm{Unif}(-5, -1)$.

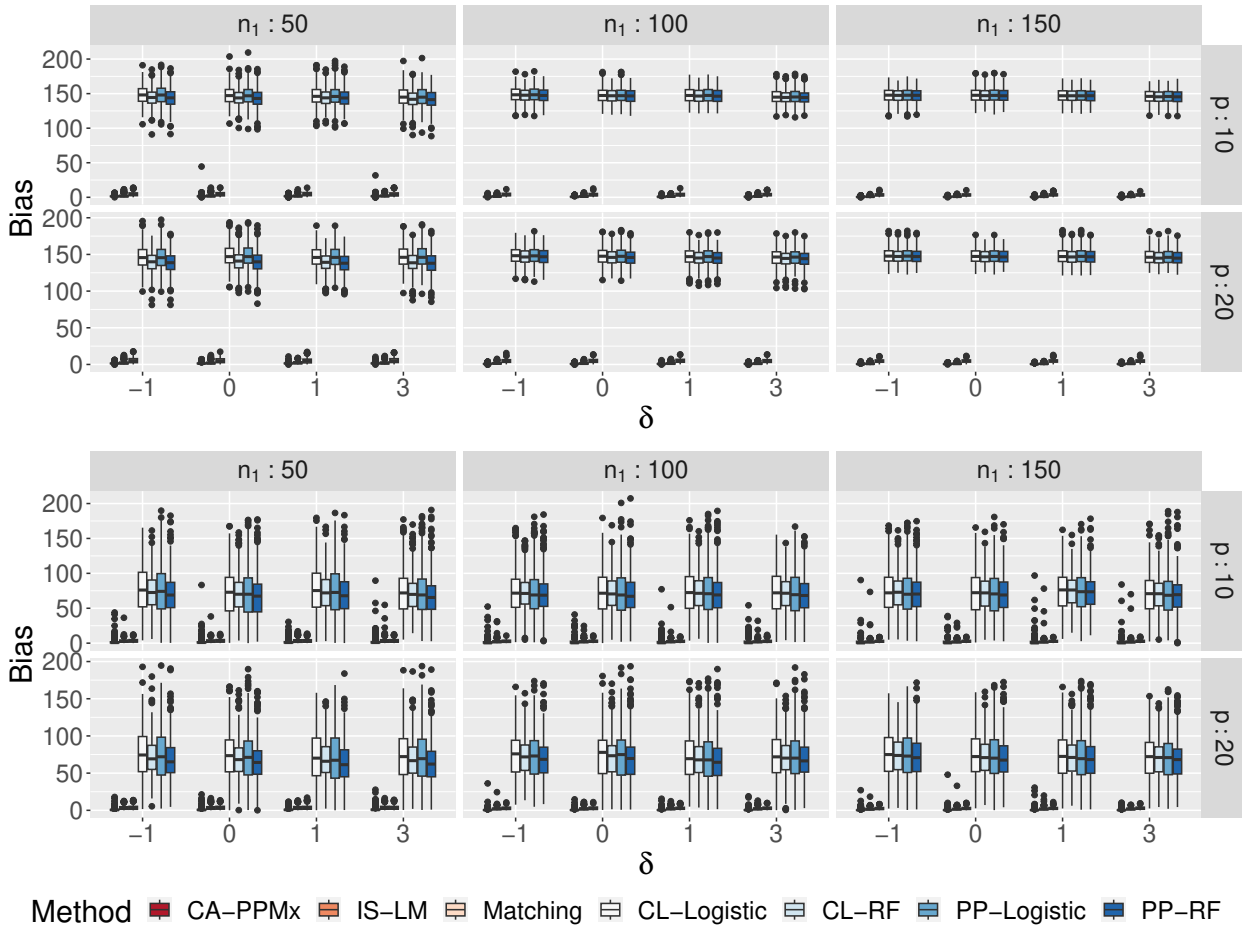
In the Interaction and Oracle scenarios we simulate the linear regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \overset{\text{iid}}{\sim} \mathrm{Unif}(-10, 10)$.

S.9.3 Implementation of Matching and PS-Based Approaches

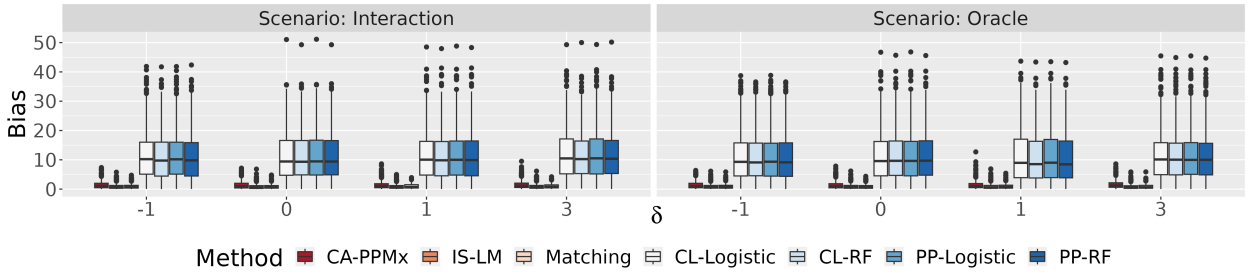
PS-based approaches: We implemented the composite likelihood and power-prior approaches using the `psrwe R` package. We set the hyperparameters as recommended in the vignette. We create 5 strata (suggested in the package vignette) and borrow n_1 patients from the RWD for all simulation studies. For the PS model, we consider both, linear logistic regression and the random forest classifier.

Matching: We implemented these approaches using the `optmatch R` package. Following the recommendations in the vignette, we set one control to be matched to each treatment. It makes the matched control population to be of the same size as the treatment arm. We then fit a linear model to estimate the treatment effect δ.

S.9.4 Bias for the Methods Considered in Section 6



(a) CAM (top) and MIX (bottom) scenarios.



(b) Interaction (left) and Oracle (right) scenarios.

Figure S.3: The bias in detecting treatment effects across different simulation setups: Seven methods are used to estimate the effects where IS-LM and CA-PPMx are based on the proposed CAM model. Panel (a) corresponds to the CAM (top) and MIX (bottom) scenarios. Panel (b) shows results under the Interaction (left side) and the Oracle (right side) scenarios.

S.9.5 Power for the Methods Considered in Section 6

The PS-based approaches yield very similar results. Therefore, for easier apprehension we only show the results for CL-RF together with the other types of methods in Tables S.1 and S.2, and the rest of the PS-based methods in Table S.3.

Table S.1: Power of detecting treatment effects under CAM and MIX scenarios

		Scenario: CAM			Scenario: MIX				Scenario: CAM			Scenario: MIX		
δ		n_1	p	Power	n_1	p	Power		n_1	p	Power	n_1	p	Power
-1	Method: CA-PPMx	50	10	0.024	50	10	0.056	Method: IS-LM	50	10	0.054	50	10	0.042
		100	10	0.032	100	10	0.066		100	10	0.080	100	10	0.090
		150	10	0.048	150	10	0.118		150	10	0.080	150	10	0.072
		50	20	0.206	50	20	0.052		50	20	0.050	50	20	0.056
		100	20	0.040	100	20	0.050		100	20	0.068	100	20	0.052
		150	20	0.062	150	20	0.030		150	20	0.118	150	20	0.064
0		50	10	0.050	50	10	0.050		50	10	0.050	50	10	0.050
		100	10	0.050	100	10	0.050		100	10	0.050	100	10	0.050
		150	10	0.050	150	10	0.050		150	10	0.050	150	10	0.050
		50	20	0.050	50	20	0.050		50	20	0.050	50	20	0.050
		100	20	0.050	100	20	0.050		100	20	0.050	100	20	0.050
		150	20	0.050	150	20	0.050		150	20	0.050	150	20	0.050
1		50	10	0.048	50	10	0.056		50	10	0.056	50	10	0.090
		100	10	0.890	100	10	0.266		100	10	0.064	100	10	0.080
		150	10	0.950	150	10	0.674		150	10	0.112	150	10	0.042
		50	20	0.058	50	20	0.142		50	20	0.064	50	20	0.048
		100	20	0.806	100	20	0.534		100	20	0.082	100	20	0.074
		150	20	0.924	150	20	0.826		150	20	0.096	150	20	0.082
3		50	10	0.866	50	10	0.056		50	10	0.146	50	10	0.132
		100	10	0.960	100	10	0.746		100	10	0.358	100	10	0.216
		150	10	0.998	150	10	0.754		150	10	0.472	150	10	0.154
		50	20	0.966	50	20	0.754		50	20	0.164	50	20	0.078
		100	20	0.958	100	20	0.900		100	20	0.312	100	20	0.228
		150	20	0.998	150	20	0.900		150	20	0.518	150	20	0.240
-1	Method: CL-RF	50	10	0.056	50	10	0.014	Method: Matching	50	10	0.050	50	10	0.076
		100	10	0.056	100	10	0.056		100	10	0.064	100	10	0.044
		150	10	0.042	150	10	0.060		150	10	0.096	150	10	0.100
		50	20	0.044	50	20	0.046		50	20	0.040	50	20	0.058
		100	20	0.076	100	20	0.034		100	20	0.046	100	20	0.076
		150	20	0.060	150	20	0.046		150	20	0.038	150	20	0.062
0		50	10	0.050	50	10	0.050		50	10	0.050	50	10	0.050
		100	10	0.050	100	10	0.050		100	10	0.050	100	10	0.050
		150	10	0.050	150	10	0.050		150	10	0.050	150	10	0.050
		50	20	0.050	50	20	0.050		50	20	0.050	50	20	0.050
		100	20	0.050	100	20	0.050		100	20	0.050	100	20	0.050
		150	20	0.050	150	20	0.050		150	20	0.050	150	20	0.050
1		50	10	0.044	50	10	0.016		50	10	0.078	50	10	0.070
		100	10	0.030	100	10	0.062		100	10	0.076	100	10	0.084
		150	10	0.040	150	10	0.042		150	10	0.150	150	10	0.106
		50	20	0.038	50	20	0.042		50	20	0.068	50	20	0.092
		100	20	0.064	100	20	0.040		100	20	0.104	100	20	0.070
		150	20	0.074	150	20	0.052		150	20	0.052	150	20	0.070
3		50	10	0.072	50	10	0.020		50	10	0.112	50	10	0.162
		100	10	0.056	100	10	0.030		100	10	0.216	100	10	0.278
		150	10	0.024	150	10	0.044		150	10	0.414	150	10	0.382
		50	20	0.034	50	20	0.066		50	20	0.154	50	20	0.132
		100	20	0.064	100	20	0.038		100	20	0.206	100	20	0.230
		150	20	0.046	150	20	0.030		150	20	0.236	150	20	0.294

Table S.2: Power of detecting treatment effects under Interaction and Oracle scenarios

Method	Scenario	δ	Power	Method	Scenario	δ	Power
CA-PPMx	Interaction	-1	0.079	IS-LM	Interaction	-1	0.085
		0	0.052			0	0.052
		1	0.047			1	0.083
		3	0.116			3	0.497
	Oracle	-1	0.077		Oracle	-1	0.091
		0	0.053			0	0.053
		1	0.084			1	0.092
		3	0.132			3	0.570
CL-RF	Interaction	-1	0.064	Matching	Interaction	-1	0.108
		0	0.053			0	0.050
		1	0.062			1	0.121
		3	0.071			3	0.575
	Oracle	-1	0.061		Oracle	-1	0.103
		0	0.053			0	0.053
		1	0.039			1	0.122
		3	0.043			3	0.752

Table S.3: Power of the PS-based methods in CAM and MIX scenarios (upper table) and Interaction and Oracle scenarios (lower table)

δ	Method: PP-Logistic	Scenario: CAM			Scenario: MIX			Method: PP-RF	Scenario: CAM			Scenario: MIX			Method: CL-Logistic	Scenario: CAM			Scenario: MIX		
		n_1	p	Power	n_1	p	Power		n_1	p	Power	n_1	p	Power		n_1	p	Power	n_1	p	Power
-1		50	10	0.082	50	10	0.022		50	10	0.062	50	10	0.022		50	10	0.068	50	10	0.016
		100	10	0.044	100	10	0.072		100	10	0.050	100	10	0.056		100	10	0.044	100	10	0.064
		150	10	0.036	150	10	0.056		150	10	0.048	150	10	0.064		150	10	0.036	150	10	0.070
		50	20	0.060	50	20	0.044		50	20	0.036	50	20	0.026		50	20	0.058	50	20	0.040
		100	20	0.062	100	20	0.038		100	20	0.074	100	20	0.036		100	20	0.060	100	20	0.034
		150	20	0.060	150	20	0.044		150	20	0.066	150	20	0.052		150	20	0.048	150	20	0.048
0		50	10	0.050	50	10	0.050		50	10	0.050	50	10	0.050		50	10	0.050	50	10	0.050
		100	10	0.050	100	10	0.050		100	10	0.050	100	10	0.050		100	10	0.050	100	10	0.050
		150	10	0.050	150	10	0.050		150	10	0.050	150	10	0.050		150	10	0.050	150	10	0.050
		50	20	0.050	50	20	0.050		50	20	0.050	50	20	0.050		50	20	0.050	50	20	0.050
		100	20	0.050	100	20	0.050		100	20	0.050	100	20	0.050		100	20	0.050	100	20	0.050
		150	20	0.050	150	20	0.050		150	20	0.050	150	20	0.050		150	20	0.050	150	20	0.050
1		50	10	0.058	50	10	0.030		50	10	0.072	50	10	0.018		50	10	0.060	50	10	0.020
		100	10	0.040	100	10	0.054		100	10	0.036	100	10	0.054		100	10	0.036	100	10	0.058
		150	10	0.028	150	10	0.042		150	10	0.044	150	10	0.034		150	10	0.030	150	10	0.048
		50	20	0.034	50	20	0.050		50	20	0.028	50	20	0.024		50	20	0.028	50	20	0.048
		100	20	0.040	100	20	0.038		100	20	0.068	100	20	0.032		100	20	0.036	100	20	0.040
		150	20	0.062	150	20	0.054		150	20	0.080	150	20	0.060		150	20	0.064	150	20	0.048
3		50	10	0.084	50	10	0.018		50	10	0.072	50	10	0.020		50	10	0.072	50	10	0.026
		100	10	0.040	100	10	0.038		100	10	0.062	100	10	0.026		100	10	0.044	100	10	0.028
		150	10	0.038	150	10	0.050		150	10	0.028	150	10	0.044		150	10	0.034	150	10	0.056
		50	20	0.052	50	20	0.064		50	20	0.038	50	20	0.042		50	20	0.048	50	20	0.064
		100	20	0.042	100	20	0.042		100	20	0.060	100	20	0.030		100	20	0.040	100	20	0.038
		150	20	0.040	150	20	0.028		150	20	0.058	150	20	0.040		150	20	0.038	150	20	0.038

Method	Scenario	δ	Power	Method	Scenario	δ	Power
CL-Logistic	Interaction	-1	0.060	PP-Logistic	Interaction	-1	0.058
		0	0.052			0	0.052
		1	0.058			1	0.058
		3	0.062			3	0.058
	Oracle	-1	0.065		Oracle	-1	0.063
		0	0.053			0	0.053
		1	0.036			1	0.036
		3	0.043			3	0.045
PP-RF	Oracle	-1	0.061	PP-RF	Interaction	-1	0.066
		0	0.053			0	0.053
		1	0.039			1	0.057
		3	0.043			3	0.064

S.9.6 Multiple Historical Controls

We consider a setup with historical controls arising from multiple sources, i.e., with $S > 2$. As mentioned earlier in Section 3.1, we merge the historical datasets and treat the merged data set as a single RWD population with increased heterogeneity. We study the performance of the CA-PPMx model in this scenario via simulation studies. We extend the MIX scenario discussed in Section 6. We generate the treatment arm $\boldsymbol{X}_{1,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \pi_{1,j} \mathrm{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$. We generate two RWD datasets from $\boldsymbol{X}_{2,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k-1} \pi_{2,j} \mathrm{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$ and $\boldsymbol{X}_{3,i} \overset{\text{iid}}{\sim} \sum_{j=2}^{k} \pi_{3,j} \mathrm{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$. In this construction, the historical populations $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are substantially different, with one distinct atom each, as well as varying weights for the common atoms. Letting $\boldsymbol{X}_{2'}$ denote the merged $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ population, we fit the CA-PPMx model on $\boldsymbol{X}_1$ and $\boldsymbol{X}_{2'}$. Note that the current trial population $\boldsymbol{X}_1$ has an extra atom compared to each of the RWD populations but the merged $\boldsymbol{X}_{2'}$ and $\boldsymbol{X}_1$ share common atoms.

We generate the response $Y_{1,i} \overset{\text{ind}}{\sim} \mathrm{N}(\delta + \boldsymbol{X}_{1,i}^{\mathrm{T}} \boldsymbol{\beta}, 1)$ and $Y_{s,i} \overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{X}_{s,i}^{\mathrm{T}} \boldsymbol{\beta}, 1)$ for $s = 2, 3$ implying δ to be the true treatment effect. We let n_2, n_2 and n_3 denote the sample sizes in the three populations, respectively where we set $n_2 = n_3 = 3 \times n_2$ in coherence with the simulation studies in Section 6. We set the dimension of the covariates $p = 10$ and repeat the the experiments for $\delta = -1, 0, 1, 3$ and $n_2 = 50, 100, 150$. We plot the power of discovering the treatment effect in Figure S.4 calculated in the exact same manner as described in Section 6. We observe that the power increases with respect to both sample size and strength of the treatment effect.

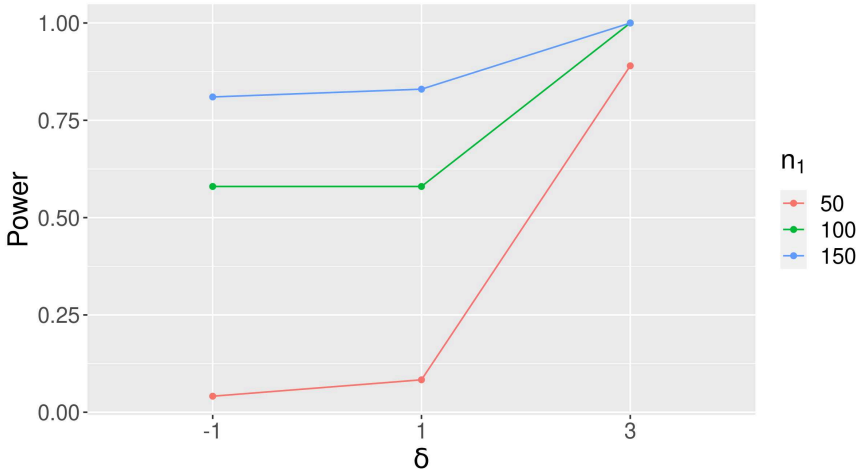


Figure S.4: Multiple historical data in the CA-PPMx model: We combine different historical datasets and combine them as a more heterogeneous single population and subsequently fit the CA-PPMx model. We observe that the power increases with respect to both sample size and strength of the treatment effect.

S.9.7 Effect of Missing Confounders

In this section we briefly study the effect of missing confounders on inference under the proposed CA-PPMx model. In particular we consider the case where a confounding factor is completely unobserved. In such cases causal inference methods are often biased; see Nguyen *et al.* (2017) and the references therein for a detailed review. However, in many applications, multivariate covariates are often correlated among each other. Several imputation methods for partially observed confounders are based on this assumption (Cole *et al.*, 2006; Moons *et al.*, 2006). In such cases, observing and using another covariate which is correlated to the missing confounder as predictor can reduce bias. We study this in a simulated example.

We consider a regression setup in a case-control study $(\boldsymbol{X}_{s,i}, Y_{s,i})$, $i = 1, \ldots, n_s$, $s = 1, 2$ with bivariate covariate $\boldsymbol{X}_{s,i} = (X_{s,i,1}, X_{s,i,2})^{\mathrm{T}}$. First, we generate $X_{s,i,1} \sim \sum_{j=1}^{k} \pi_{s,j} \mathrm{N}(\mu_j, 0.01)$ and subsequently generate $X_{s,i,2} = mX_{s,i,1} + \varepsilon_{s,i}$ where $\varepsilon_{s,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$ and $m \in \mathbb{R}$. Then, we generate the responses $Y_{1,i} = \delta + \beta X_{1,i,1} + \epsilon_{1,i}$ and $Y_{2,i} = \beta X_{2,i,1} + \epsilon_{2,i}$ where $\epsilon_{s,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$ implying δ to be the true treatment effect. Thus conditionally on the $X_{s,i,1}$'s, the responses $Y_{s,i}$'s are independent of the $X_{s,i,2}$'s. We take $n_2 = 50$, $n_2 = 300$, $k = 3$, $(\mu_2, \mu_2, \mu_3) = (-3, 0, 3)$, $\delta = 3$ and $\beta = 1$. We repeat the simulation experiment independently 100 times and randomly generate the $\pi_{s,j}$'s in each replicate.

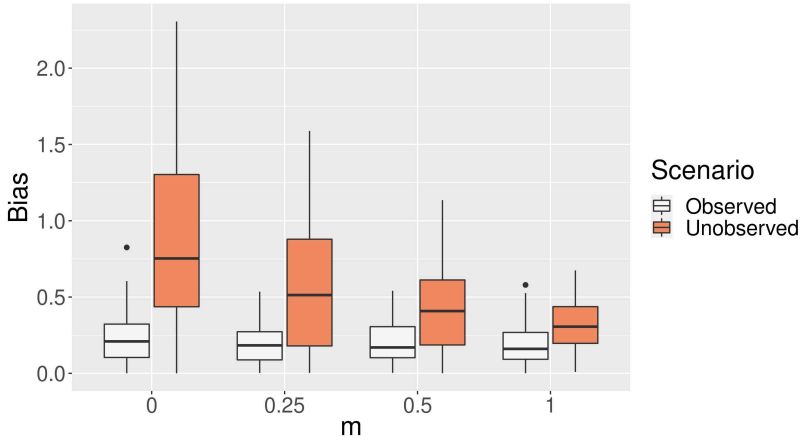


Figure S.5: Effect of missing confounder in the CA-PPMx model: The bias in estimating the treatment effect decreases as the correlation between the observed covariate and the unobserved confounder increases.

We consider two analysis scenarios: (1) *Unobserved*: $X_{s,i,1}$ is assumed to be unobserved and the CA-PPMx model is fitted using $(X_{s,i,2}, Y_{s,i})$; (2) *Observed*: the CA-PPMx model is fitted using $(\boldsymbol{X}_{s,i}, Y_{s,i})$. We compute the bias in estimating the treatment effect δ for varying values of m in both scenarios. We show boxplots of the biases over the repeat simulations in Figure S.5.

Note that for $m = 0$, $X_{s,i,1}$ and $X_{s,i,2}$ are uncorrelated. Additionally, $|\mathrm{corr}(X_{s,i,1}, X_{s,i,2})|$ is an increasing function of $|m|$. Coherently, the bias is maximum in the *Unobserved* scenario

for $m = 0$ as the $X_{s,i,2}$'s carry no information regarding the confounding factor $X_{s,i,1}$'s. The marginal correlation between the observed covariate and the response increases with m and accordingly we see a reduction in the bias. This simulation study indicates that the CA-PPMx method will not yield terribly biased results as long as the data includes observed covariates that are correlated to the unmeasured confounder.

S.9.8 Computation Times for the CA-PPMx Method

In this section we report computation times of the MCMC sampler proposed in Section S.8 across different sample sizes and covariate dimensions. We consider the CAM and MIX scenarios and the exact same simulation setups discussed in Section 6 of the main paper. Since the model implementation times do not depend on the treatment effect size, we report the computation times for $\delta = 3$ only. Computation times for $6,000$ MCMC iterations in seconds for a single repeat simulation on an Intel Core i9-13900K CPU with 128GB of RAM are provided in Figure S.6 where we see that the computational cost increases with the covariate dimension p as well as the sample size n_1.

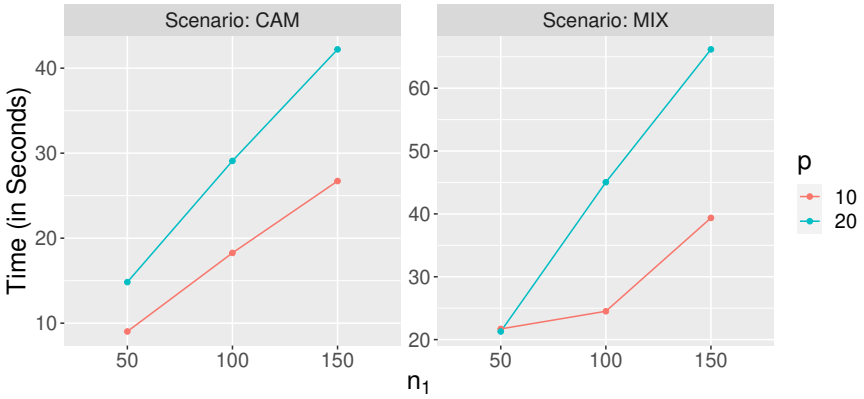


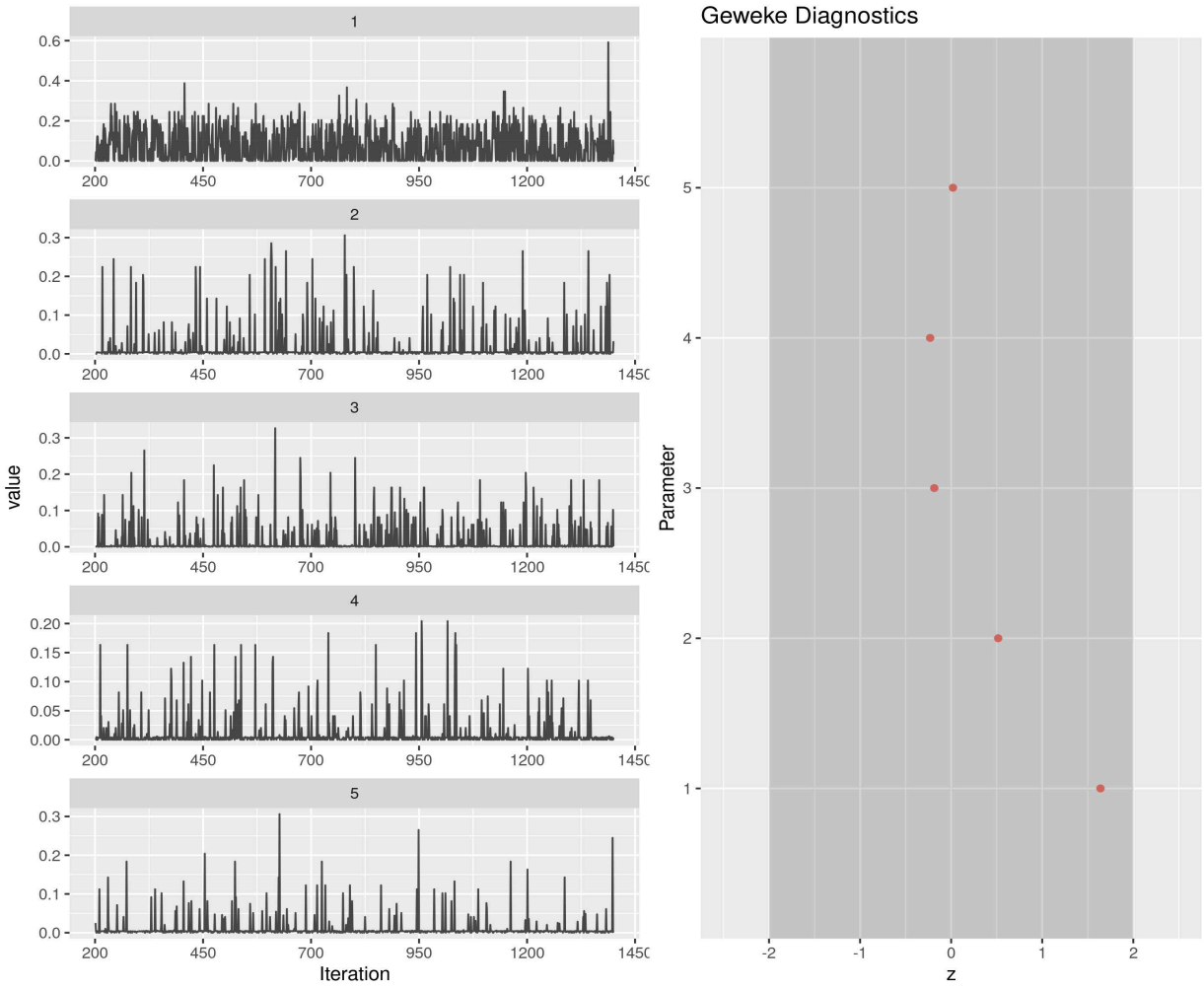
Figure S.6: Computation times of the MCMC sampler in seconds: n_1 and p denotes the number of patients in the current trial arm and the dimension of the covariates, respectively.

S.10 MCMC Diagnostics

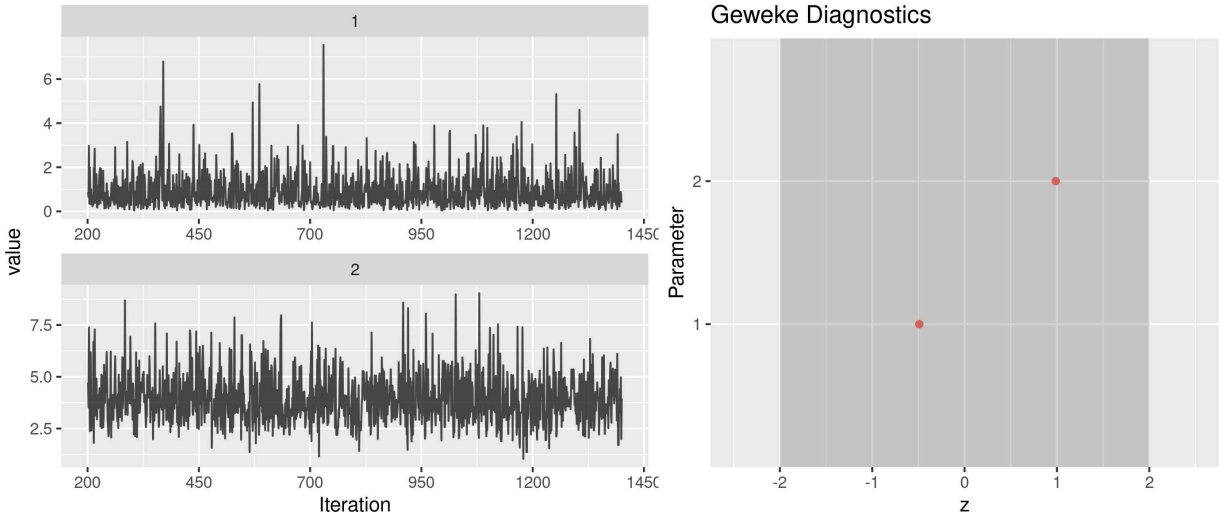
In this section, we provide some convergence diagnostics of the MCMC sampler discussed in Section S.8 for one trial replicate discussed in Section 7. We show traceplots and Geweke's convergence diagnostics (Geweke, 1992) for some selected parameters, using an implementation in the `ggmcmc` `R` package (Fernández-i Marín, 2016).

Recall the importance resampling weights $w_i \propto \frac{\pi_{1,c_{2,i}}}{n_{2,c_{2,i}}}$ in Eqn (5) attached to the historical patients. We evaluate MCMC convergence diagnostics for the five w_i's with the largest posterior means, the lognormal hyperparameters μ_0 and b_0 mentioned in Step 3 and the

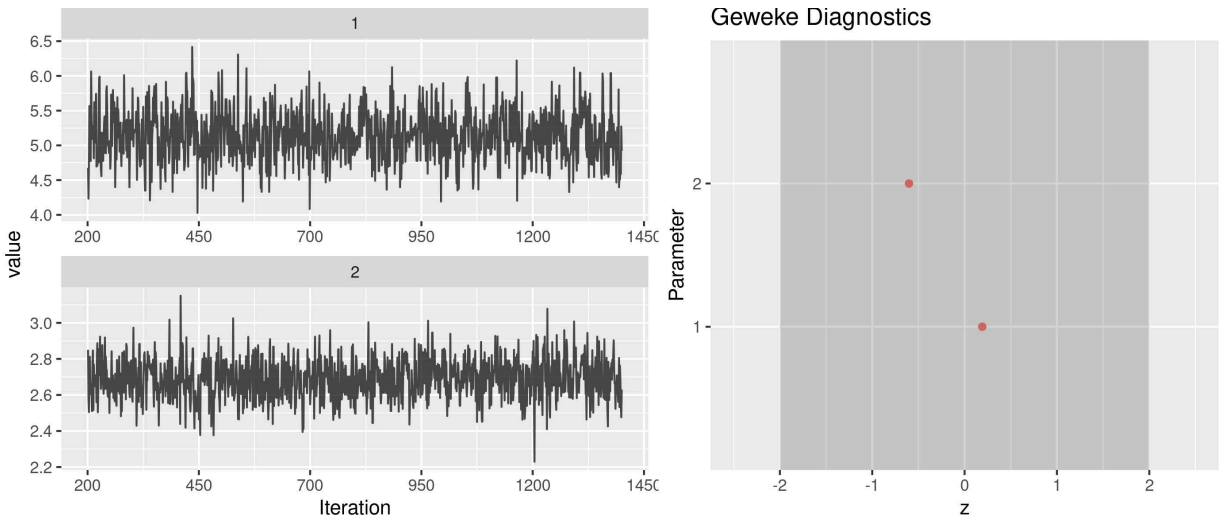
Dirichlet mixture hyperparameters α_1 and α_2 in Step 5 of the MCMC sampler in Section S.8. The results, provided in Figure S.7, do not suggest any convergence or mixing issues.



(a) MCMC analysis of the largest 5 importance resampling weights.



(b) MCMC analysis of the Dirichlet mixture hyperparameters α_2 and α_2.



(c) MCMC analysis of the lognormal hyperparameters μ_0 and b_0.

Figure S.7: MCMC convergences diagnostics for some selected parameters: Panel (a), (b) and (c) shows results for the top five $w_i \propto \frac{\pi_{1,c_{2,i}}}{n_{2,c_{2,i}}}$ with largest posterior means, the lognormal hyperparameters μ_0 and b_0 and the Dirichlet mixture hyperparameters α_1 and α_2 in Step 5, respectively. In each panel, we show the corresponding traceplots across the thinned out MCMC samples on the left, and Geweke's diagnostics on the right.

References

Cao, J., Moosman, A., and Johnson, V. E. (2010). A Bayesian Chi-squared goodness-of-fit test for censored data models. *Biometrics*, **66**, 426–434.

Chen, W.-C., Wang, C., Li, H., Lu, N., Tiwari, R., Xu, Y., and Yue, L. Q. (2020). Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*, **30**, 508–520.

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, **35**, 1074–1081.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, **195**, 216–222.

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

Fernández-i Marín, X. (2016). ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, **70**, 1–20.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, **4**, 641–649.

Hey, S. P. and Kimmelman, J. (2014). The questionable use of unequal allocation in confirmatory trials. *Neurology*, **82**, 77–79.

Ishwaran, H. and Zarepour, M. (2002a). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, **12**, 941–963.

Ishwaran, H. and Zarepour, M. (2002b). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 269–283.

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis*, **2**, 719–733.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, **27**, 435–454.

Lu, N., Wang, C., Chen, W.-C., Li, H., Song, C., Tiwari, R., Xu, Y., and Yue, L. Q. (2022). Leverage multiple real-world data sources in single-arm medical device clinical studies. *Journal of Biopharmaceutical Statistics*, **32**, 107–123.

Meloun, M. and Militký, J. (2011). The exploratory and confirmatory analysis of univariate data. In *Statistical Data Analysis*, pages 25–71. Woodhead Publishing India.

Moons, K. G., Donders, R. A., *et al.* (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, **59**, 1092–1101.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, **20**, 260–278.

Nabors, L. B., Fink, K. L., *et al.* (2015). Two cilengitide regimens in combination with standard treatment for patients with newly diagnosed glioblastoma and unmethylated MGMT gene promoter: Results of the open-label, controlled, randomized phase II CORE study. *Neuro-Oncology*, **17**, 708–717.

Nguyen, T.-L., Collins, G. S., *et al.* (2017). Magnitude and direction of missing confounders had different consequences on treatment effect estimation in propensity score analysis. *Journal of Clinical Epidemiology*, **87**, 87–97.

Page, G. L., Quintana, F. A., and Müller, P. (2022). Clustering and prediction with variable dimension covariates. *Journal of Computational and Graphical Statistics*, **31**, 466–476.

Robert, C. P. and Roberts, G. (2021). Rao–Blackwellisation in the Markov chain Monte Carlo era. *International Statistical Review*, **89**, 237–249.

Skare, O., Bølviken, E., and Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, **30**, 719–737.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25**, 1–21.

Stupp, R., Hegi, M. E., *et al.* (2014). Cilengitide combined with standard treatment for patients with newly diagnosed glioblastoma with methylated MGMT promoter (CENTRIC EORTC 26071-22072 study): A multicentre, randomised, open-label, phase 3 trial. *The Lancet Oncology*, **15**, 1100–1108.

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.

Wang, C., Li, H., Chen, W.-C., *et al.* (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, **29**, 731–748.

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.