Supplementary Materials for

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

Noirrit Kiran Chandra[a] (noirritchandra@gmail.com)
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.

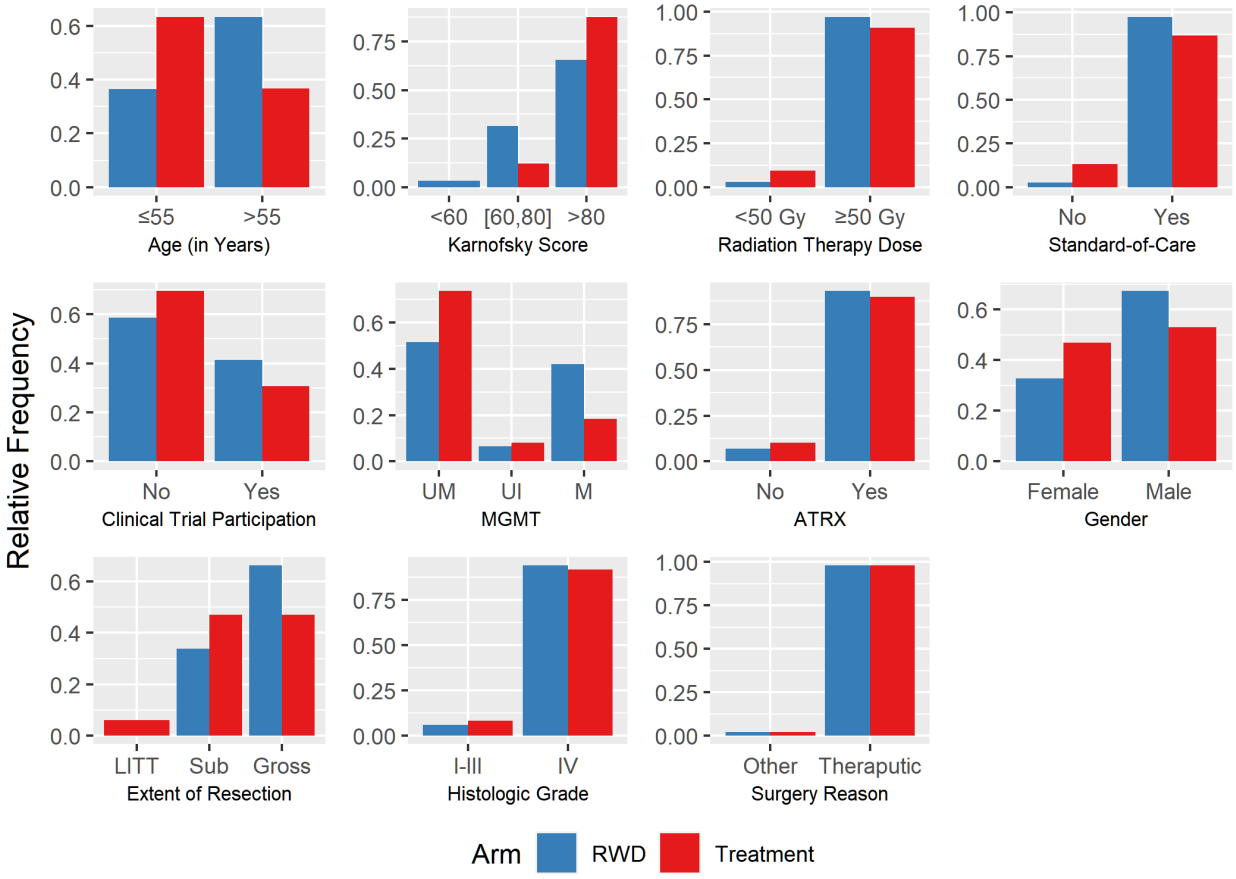


Figure S.1: Relative frequency plots of the covariates in the two treatment arms.

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 $\text{Unif}(0, 1)$ 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 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

$$\textbf{likelihood: } Y_i \mid \boldsymbol{X}_i \overset{\text{ind}}{\sim} \text{N}(\alpha + \boldsymbol{\beta}^{\text{T}} \boldsymbol{X}_i, \sigma^2);$$

$$\textbf{prior: } (\alpha, \boldsymbol{\beta}) \sim \text{N}_{p+1}(\boldsymbol{0}, 10 \times \boldsymbol{I}_{p+1}),\ \sigma^{-2} \sim \text{Ga}(0.1, 0.1). \tag{S.7}$$

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.

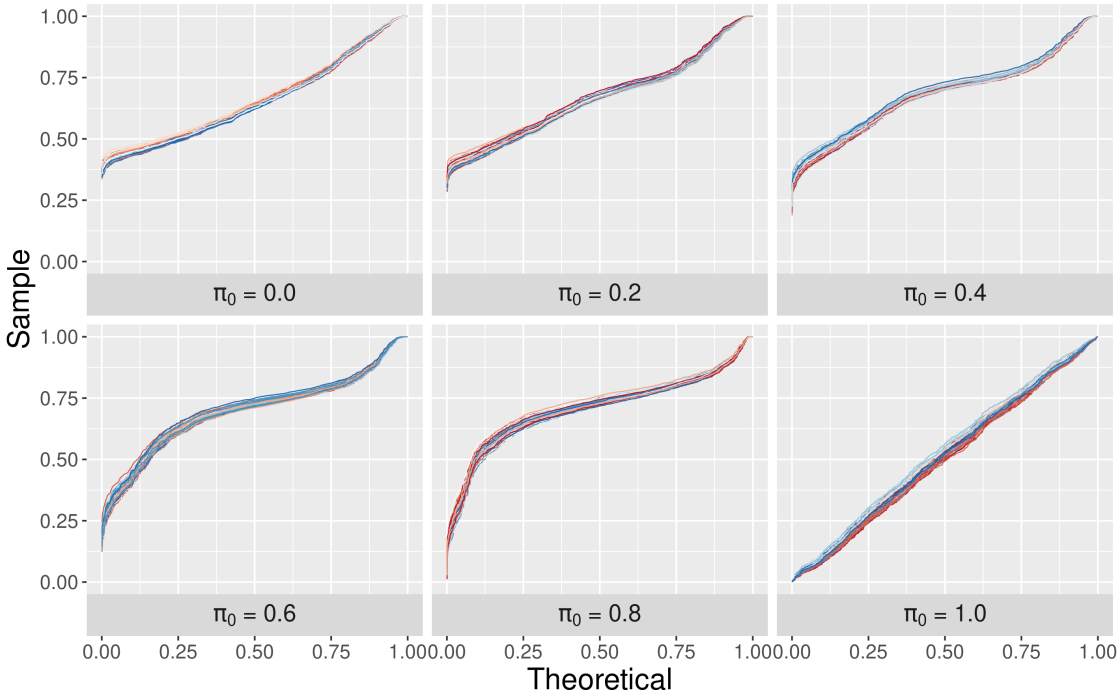


Figure S.2: Quantile-quantile plots for increasing model misspecification: Data are generated from model (S.6) for different values of π_0 and the Bayesian linear regression model in Eqn (S.7) is fitted where $\pi_0 = 0.0$ and 1.0 denote the extreme misspecified model and the true model, respectively.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

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

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 $\beta_4 \sim \mathrm{Unif}(-5, -1)$.

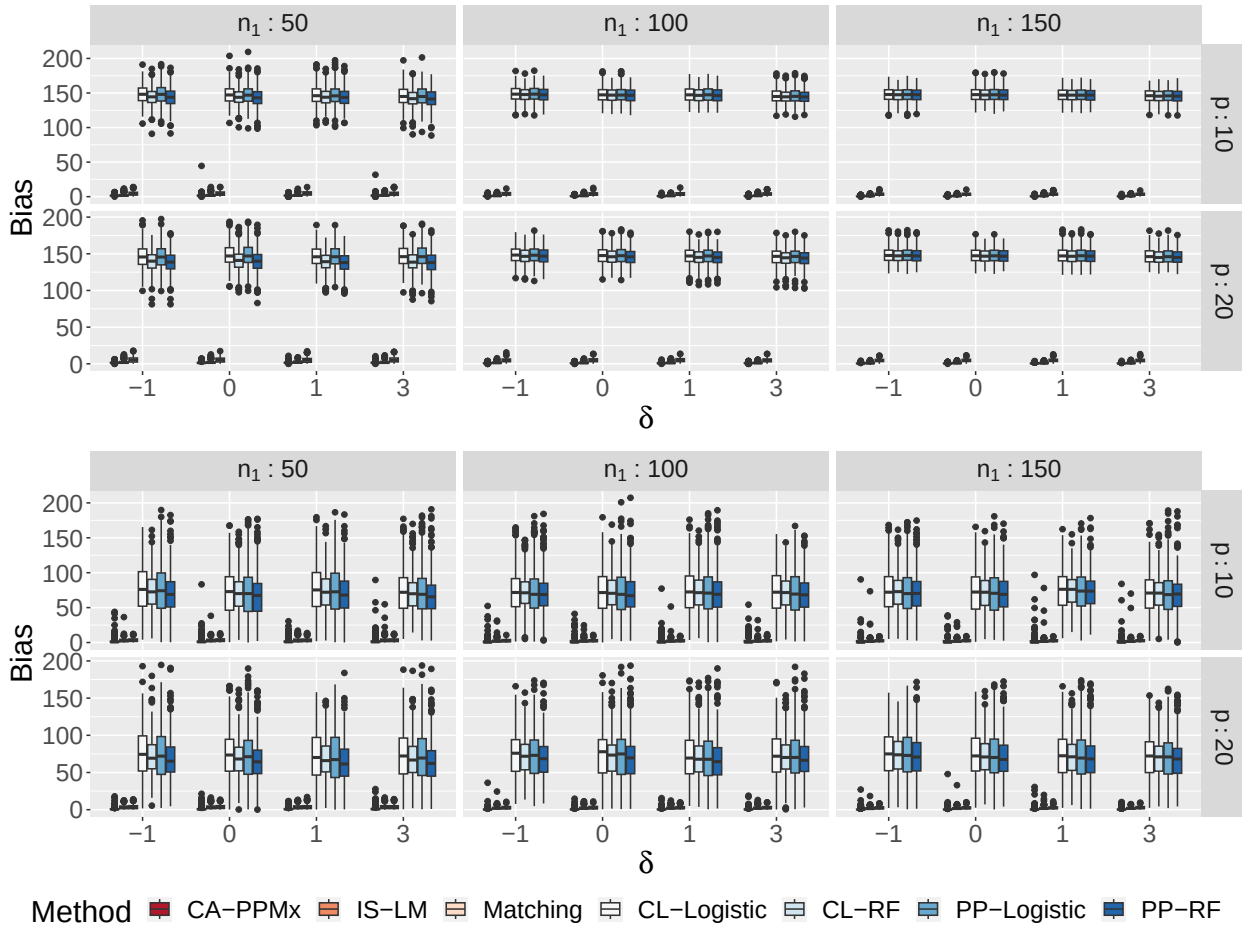
In the Interaction and Oracle scenarios we simulate the linear regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \overset{\text{iid}}{\sim} \mathrm{Unif}(-10, 10)$.

S.9.3 Implementation of Matching and PS-Based Approaches

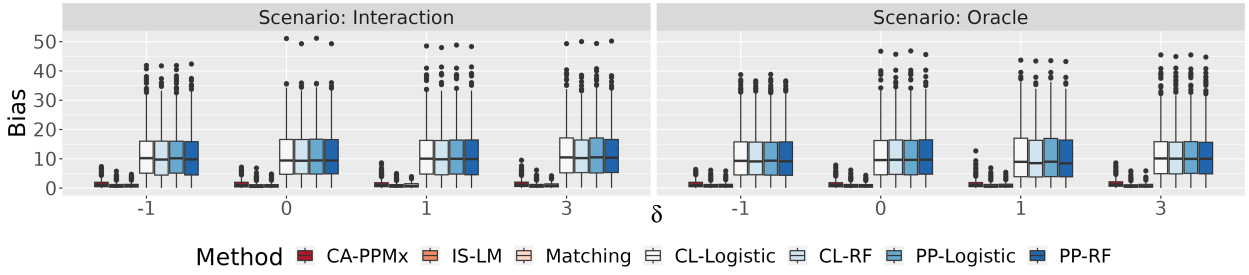
PS-based approaches: We implemented the composite likelihood and power-prior approaches using the `psrwe R` package. We set the hyperparameters as recommended in the vignette. We create 5 strata (suggested in the package vignette) and borrow n_1 patients from the RWD for all simulation studies. For the 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

δ		Scenario: CAM			Scenario: MIX				Scenario: CAM			Scenario: MIX		
		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 $\boldsymbol{X}_{1,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \pi_{1,j} \text{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$. We generate two RWD datasets from $\boldsymbol{X}_{2,i} \overset{\text{iid}}{\sim} \sum_{j=1}^{k-1} \pi_{2,j} \text{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$ and $\boldsymbol{X}_{3,i} \overset{\text{iid}}{\sim} \sum_{j=2}^{k} \pi_{3,j} \text{N}_p(\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}_p)$. 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 $Y_{1,i} \overset{\text{ind}}{\sim} \text{N}(\delta + \boldsymbol{X}_{1,i}^{\text{T}} \boldsymbol{\beta}, 1)$ and $Y_{s,i} \overset{\text{ind}}{\sim} \text{N}(\boldsymbol{X}_{s,i}^{\text{T}} \boldsymbol{\beta}, 1)$ 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.

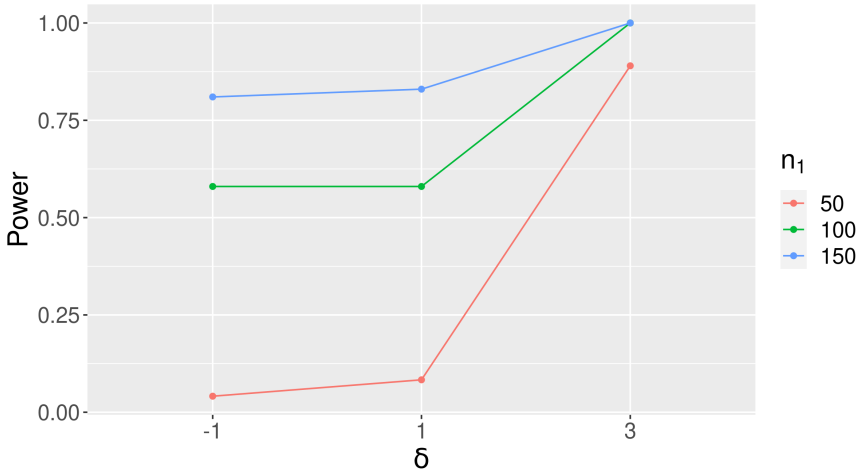


Figure S.4: Multiple historical data in the CA-PPMx model: We combine different historical datasets and combine them as a more heterogeneous single population and subsequently fit the CA-PPMx model. We observe that the power increases with respect to both sample size and strength of the treatment effect.

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.

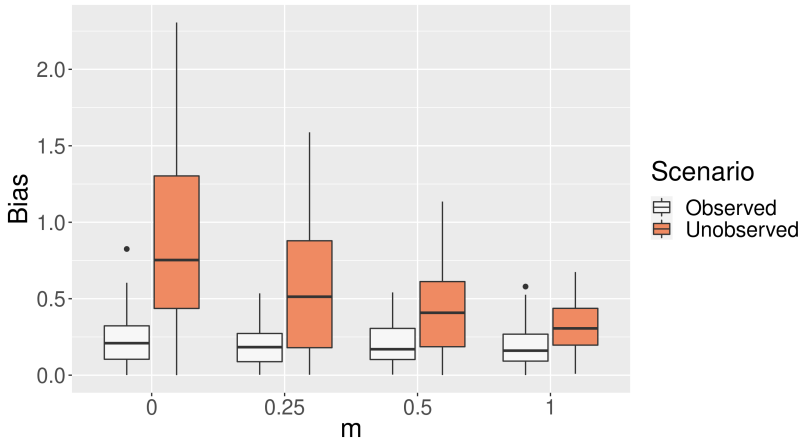


Figure S.5: Effect of missing confounder in the CA-PPMx model: The bias in estimating the treatment effect decreases as the correlation between the observed covariate and the unobserved confounder increases.

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.

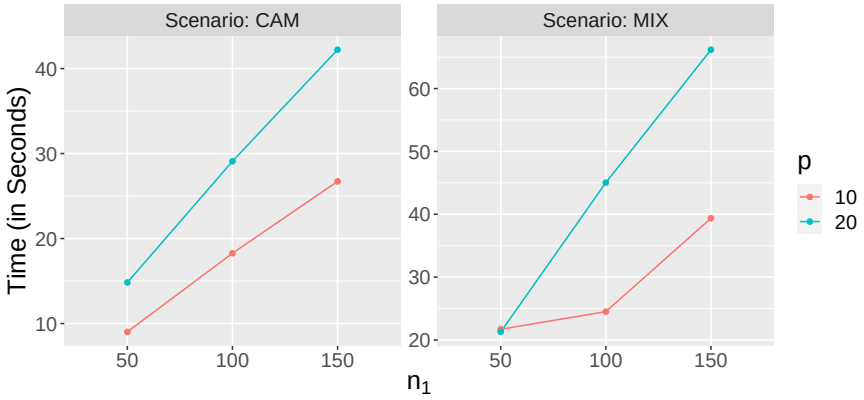


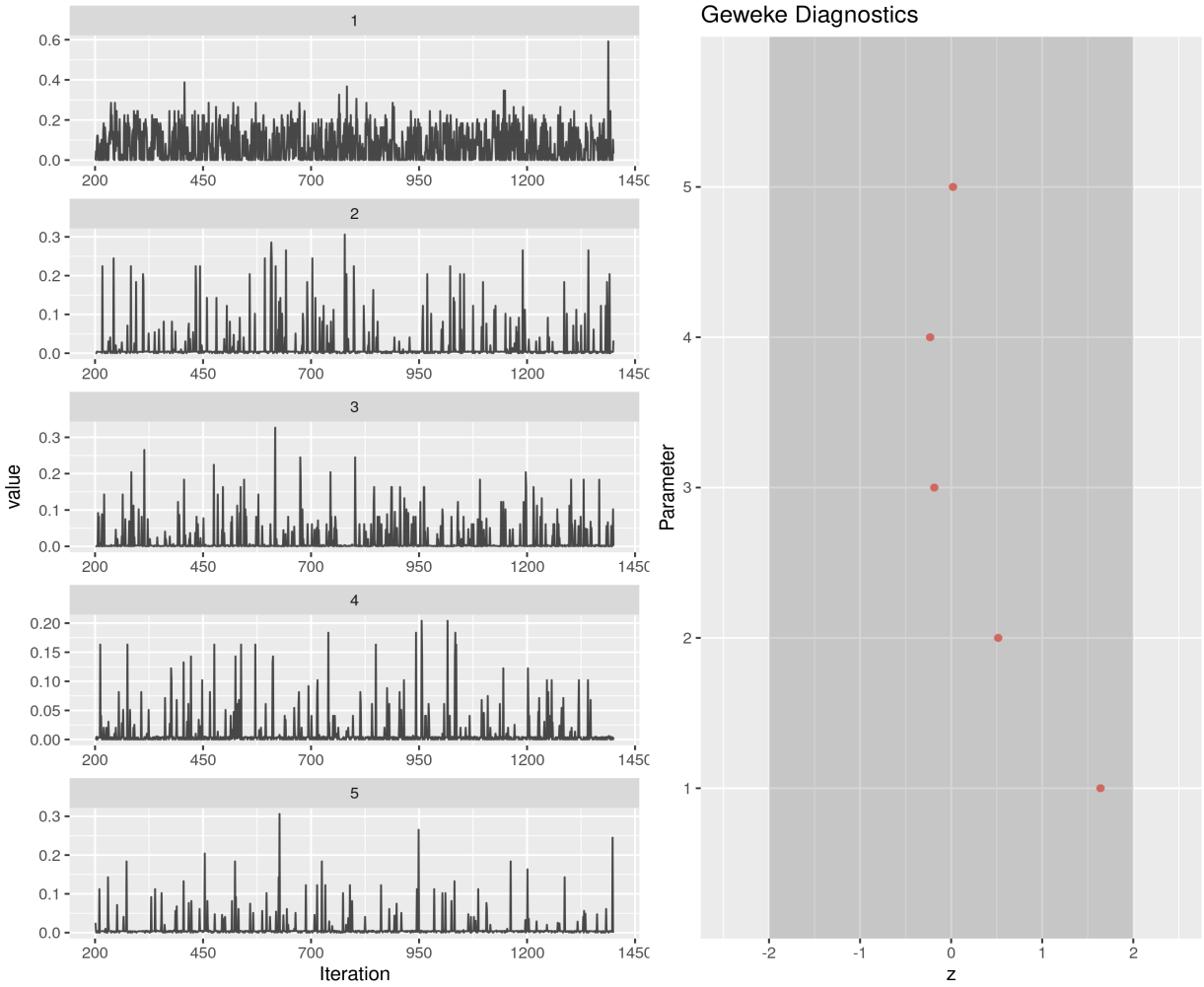
Figure S.6: Computation times of the MCMC sampler in seconds: n_1 and p denotes the number of patients in the current trial arm and the dimension of the covariates, respectively.

S.10 MCMC Diagnostics

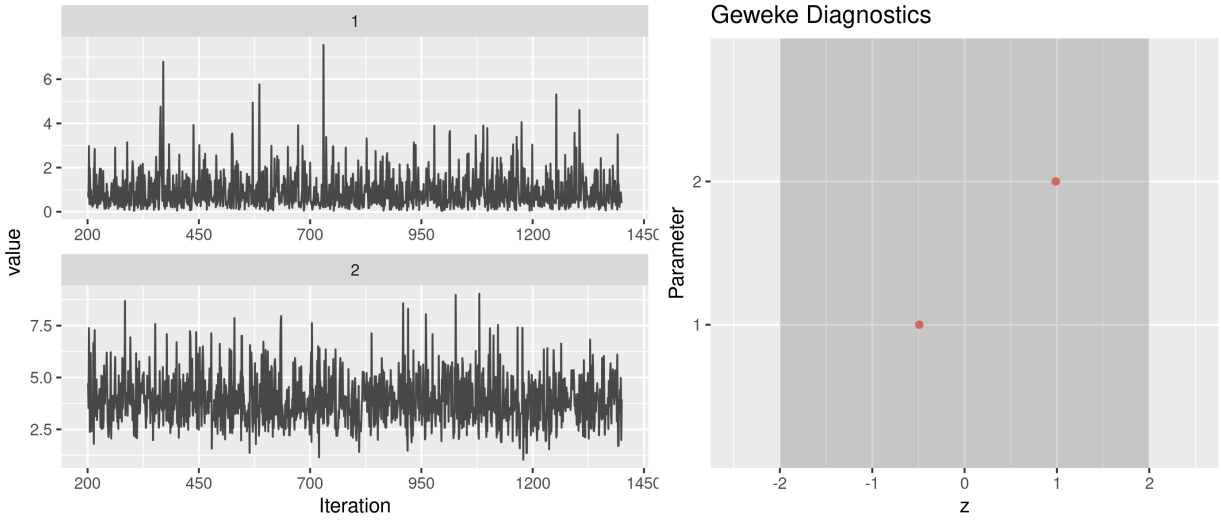
In this section, we provide some convergence diagnostics of the MCMC sampler discussed in Section S.8 for one trial replicate discussed in Section 7. We show traceplots and Geweke's convergence diagnostics (Geweke, 1992) for some selected parameters, using an implementation in the `ggmcmc` `R` package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

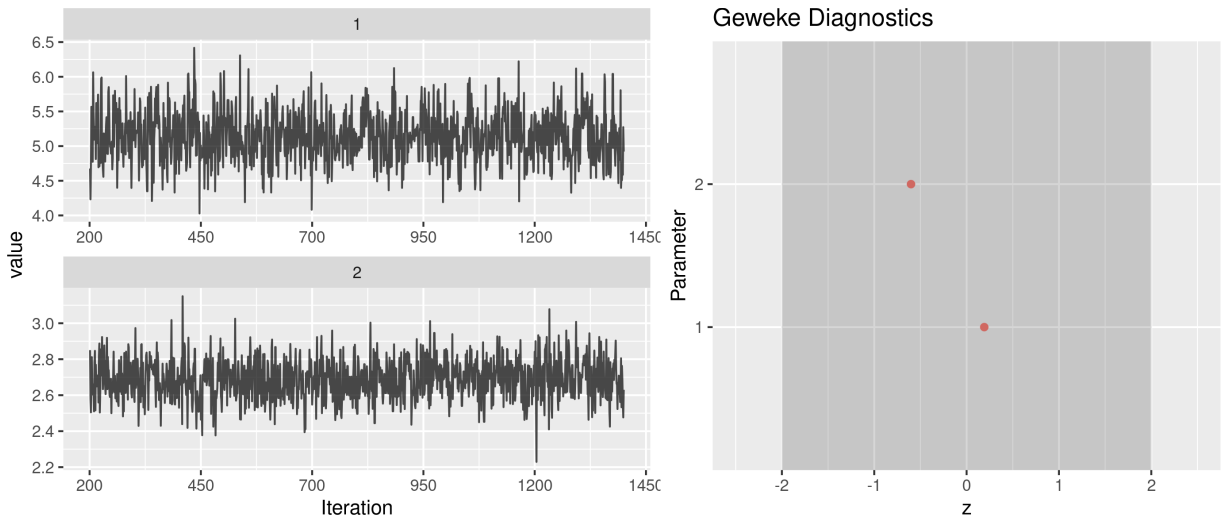
Dirichlet mixture hyperparameters α_1 and α_2 in Step 5 of the MCMC sampler in Section S.8. The results, provided in Figure S.7, do not suggest any convergence or mixing issues.



(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.