Biostatistics, 2026, **27**(1), kxag003
https://doi.org/10.1093/biostatistics/kxag003
Article

SAM-HC: a Bayesian nonparametric construction of hybrid control for randomized clinical trials using external data

Dehua Bi[1], Tianjian Zhou[2], Wei Zhong[3,], and Yuan Ji[1,*]

[1]Department of Public Health Sciences, The University of Chicago, IL, United States
[2]Department of Statistics, Colorado State University, CO, United States
[3]Oncology Biostatistics, Pfizer Inc, United States

*Corresponding author: Department of Public Health Sciences, The University of Chicago, 5841 S Maryland Ave. MC2000, Chicago, IL 60637, United States. Email: yji@bsd.uchicago.edu

Summary

In situations where high-quality external data are available or when it is challenging to recruit participants to the control arm of a randomized and controlled clinical trial (eg rare or pediatric diseases), it is desirable to borrow information from external data to augment the control arm. However, a main challenge in borrowing information from external data is to accommodate potential heterogeneous subpopulations across the external and trial data. We apply a Bayesian nonparametric model called the Shared Atoms Model (SAM) to identify overlapping and unique subpopulations across datasets, with which we restrict the information borrowing to the common subpopulations. This forms a hybrid control (HC) that leads to more precise estimation of treatment effects. The degree of information borrowing is confined by the sample size and degree of similarity in outcomes. Simulation studies demonstrate the robustness of the new method, and an application to an Atopic Dermatitis dataset shows improved treatment effect estimation.

Keywords dependent clustering, overlapping clusters, power priors, real world data, shared atoms model

1. Introduction

Randomized controlled trials (RCTs) are the gold standard to objectively assess the treatment effect of a new drug over a control. Equal randomization in RCTs is common in practice due to its ease of implementation and the perceived benefits for statistical inference (Hey and Kimmelman 2014). Nevertheless, enrolling patients under a design with equal randomization can sometimes be challenging, like in rare diseases, pediatric trials, or settings where an r:1 ($r > 1$) randomization ratio is used to enhance patient enrollment. With the availability of historical trial data or real-world data (RWD) like the electronic health records, statistical models have been proposed to borrow

Received: October 15, 2023. **Revised**: November 2, 2025. **Accepted**: November 17, 2025

information in these data to estimate treatment effects more accurately. For example, when a standard of care has been widely tested or administered in a patient population, the available response data could be used to augment the control arm in a clinical trial and form a hybrid control (HC). Due to the augmented information in the HC, a more precise estimation of the treatment effect can be achieved when the outcome distributions between the current control and external data are well aligned.

In drug development, the US FDA has provided a "Framework for FDA's Real-World Evidence Program" (US FDA 2018) on using external data to support the approval of new indications for drugs or to meet post-approval study requirements. More recently, a guidance on the design and conduct of external controlled trials for Drug and Biological products (FDA, US 2023) has been released. The document emphasizes the importance of ensuring that the trial eligibility criteria can be applied to the external control arm in order to obtain a population comparable to that of the clinical trial. Thus, it is critical to ensure the similarity of patient baseline characteristics between the external data and the current RCT. Another issue discussed is the extent to which one may borrow information (Chen et al. 2020). Historical trial data and RWD often have substantially larger sample sizes than the current RCT data, so caution is needed to ensure that the borrowed information does not overly influence the results of the current trial. For example, one might quantify the effective sample size (ESS) of the information borrowed and require that ESS is no larger than the actual size of the control arm in the RCT. Therefore, oftentimes information from external data is discounted to avoid overwhelming the statistical inference of the current study.

In the literature, many statistical methods have been proposed to borrow information from external data for RCTs. Bayesian models, such as the power prior (PP) (Ibrahim and Chen 2000), commensurate prior (CP) (Hobbs et al. 2012), and the robust meta-analytic-predictive prior (RMAP) (Schmidli et al. 2014) construct hierarchical models for external and current trial data based on outcomes alone. Specifically, the PP method assumes that the treatment outcome parameters are identical between the current trial and external data. It utilizes a discounting factor to reflect the user's prior belief regarding the similarity between the historical data and the current trial, thereby discounting the likelihood of the historical data. CP uses different parameters for the current trial and the external data, but assuming the parameters of the trial follow a prior distribution with a mean equal to the parameters for the external data. RMAP employs a mixture of a meta-analytic-predictive (MAP) prior, which is an informative prior, and a vague prior (robust component) to mitigate the potential issue of over-borrowing. Additionally, several other methods such as the BaSe-MAP prior (Hupf et al. 2021) and the BEATS design (Bi et al. 2023) also rely solely on outcomes for data borrowing.

Another class of methods uses propensity scores (PS) to identify matched patients between external data and the current trial data. Unlike outcome-based approaches, these methods select subpopulations from the external data based on covariates, aiming to ensure similar baseline characteristics between patients in the current trial and those borrowed from the external data. For instance, Chen et al. (2020) proposed the propensity-score integrated composite likelihood (PSCL) method to address situations where only a subset of patients in the external data are comparable to those in the current trial. Similarly, Lu et al. (2022) introduced the PS-PP approach, Liu et al. (2021) proposed the PS-MAP prior, and Baron et al. (2022) developed the Bayesian divide-and-conquer PS-based approaches for single arm trials.

Recently, Zhou and Ji (2021) applies the BART (Chipman et al. 2010), a tree-based method that incorporates different sources of data and adjust covariates in the estimation of treatment effects. Alt et al. (2024) introduced the LEAP prior, which dynamically borrows information from historical data based on both outcomes and covariates, assuming that a subset of individuals in the historical data are exchangeable with those in the current RCT. Another recent study by Chandra et al. (2023) utilizes Bayesian nonparametric models (BNP) to identify "common clusters" of patients across the current trial and the external data. The BNP model has the ability to automatically cluster patients in the current trial and the external data based on baseline covariates. Their method, called CA-PPMx, assumes the external data consists of all the subpopulations that are present in the current trial. However, neither LEAP nor CA-PPMx accounts for potential scenarios where the

current trial includes clusters, or subgroups of patients, that are not represented in the historical or external data.

In this work, we also aim to identify subpopulations within the external data to facilitate information borrowing based on patient-level covariates. However, we do not use PS for this purpose. As noted by Zhao (2004), King and Nielsen (2019), and Chandra et al. (2023), matching patients based on their PS does not necessarily imply matching of covariates. In addition, PS-based methods are often sensitive to model specification for estimating the PS.

Motivated by CA-PPMx, we propose a BNP approach, called SAM-HC, for constructing an HC arm for an RCT using external data. Here, SAM refers to the Shared Atoms Model, which is a BNP model in Bi and Ji (2023) that generates overlapping clusters. We assume that the current RCT and external data may share common subpopulations of patients, while each may consist of unique ones as well (ie there could be subpopulations in the RCT that is not part of the historical data). In other words, SAM can identify both common and unique subpopulations across observations arranged in groups. Using SAM, the HC is constructed by only borrowing information from the common subpopulations between the external data and the control in the RCT. We employ a power prior to discount the information borrowing based on outcome distributions. In addition, the BNP models in the proposed SAM-HC method generate random clusters characterized by a posterior distribution. Therefore, the entire statistical inference is model based and variabilities on the clustering and treatment effect estimates are properly accounted for.

The remainder of the paper is constructed as follows. We first provide a general overview of the method in Section 2. Then, we review the SAM method in Section 3. We provide a detailed description of the proposed SAM-HC method in Section 4. We present the simulation setup and results in Section 5 comparing our method to the PSCL method, the PS-MAP approach, and a baseline method that does not involve information borrowing. In Section 6, we apply SAM-HC to real-life trial data. Finally, we conclude our work in Section 7.

2. General methodology

We consider an ongoing RCT with an $r:1$ ($r > 1$) randomization ratio between the treatment and control arms. There also exists an external dataset of patients with the same disease who received the same control regimen as in the current RCT. For example, both the RCT control arm and the external data may involve standard chemotherapy, while the RCT treatment arm receives a new immunotherapy. We denote the RCT treatment arm as group 1 ($j = 1$), the RCT control arm as group 2 ($j = 2$), and the external data as group 3 ($j = 3$). Let n_j be the number of patients in group j for $j = 1, 2, 3$. Then $N = n_1 + n_2$ is the RCT sample size, and $n_1/n_2 \approx r$ by design. We denote $y_j = \{y_{i,j}; i = 1, \ldots, n_j\}$ the patients outcome data in group j, and $x_{i,j} = \{x_{i,j,1}, \ldots, x_{i,j,p}\}$ a p-dimensional covariate vector of patient i's baseline covariates in group j.

To infer the treatment effect, we make the following assumptions in addition to the implied stable unit treatment value assumption (SUTVA). First, we assume that heterogeneous subpopulations of patients are present in both the trial and the external data, defined by the distributions of baseline covariates. We use the terms "cluster" and "subpopulation" interchangeably. Subpopulations that appear in multiple groups are called "common clusters," whereas a subpopulation present in only 1 group is a "unique cluster." Because treatment assignment in the RCT is randomized and patients adhere to their assigned arm, an unbiased estimate of the within-subpopulation treatment effect is attainable for treatment and control patients in the same cluster. Second, we assume within-cluster exchangeability: conditional on cluster membership, outcomes are exchangeable for patients in the current control and in the external data. Third, we assume conditional independence of outcomes and covariates given cluster membership; therefore, once cluster membership is learned, no additional residual covariate adjustment is required when borrowing information.

Based on these assumptions, statistical inference can proceed with the following 2 simple steps: Step 1, identify latent cluster membership using baseline covariates; Step 2, dynamically borrow information from the external data to estimate cluster-specific treatment effects. Because the joint distribution of covariates is difficult to specify parametrically and the number of clusters is typically unknown, we employ a BNP model—the Shared Atoms Model (SAM)—to infer cluster

Table 1 Notations used for the proposed method.

Symbol	Description
i, j, k	Index of the patients, group, and cluster.
r	Randomization ratio in the current trial.
N, I	Sample size in current trial and. desired maximum sample size after information borrowing.
$y_{i,j}, x_{i,j}$	outcomes and covariates of patient i in group j.
$Z_{i,j}, Z$	Cluster membership of patient i in group j and the membership matrix for all subjects.
$\pi_{j,k}$	Cluster k's size/proportion in group j.
$\pi^*_{2,k}$	Computed proportion to be borrowed into cluster k from the external data.
$A_{j,k}$	Index of the patients in cluster k from group j.
$n_{j,k}$	Number of patients in cluster k from group j.
O_j	Index of the clusters in group j.
$C_{j,j'}$	Index of the clusters that is common between 2 different groups: j and j'.
$\theta_{1,k}, \theta_{2,k}$	Cluster k's response in treatment (group 1) and control (group 2) arms.
ρ_k	Overlapping coefficient of cluster k's outcomes between current-control and external data.
α_k	Power parameter used to borrow information from cluster k in external data.
$\Delta_k, \Delta(Z)$	Cluster k specific treatment effect and overall treatment effect based on membership matrix Z.

membership using baseline covariates (Bi and Ji 2023). After learning cluster membership, we borrow information from the external data via a power-prior approach, with the borrowing weight (power parameter) adaptively tuned to the similarity of the outcome distributions between the current-control and the external data. A notation summary table is provided in Table 1, and a complete list appears in Table 1, supplementary material.

3. Review shared atoms model (SAM)

We apply the SAM in Bi and Ji (2023) for modeling the subpopluations of patients across groups. An example of the clustering structures from SAM is shown in Fig. 1. In this illustration, each color represents a cluster. The blue and green clusters are common and the purple cluster is unique to group 3. If one knows the clustering pattern in Fig. 1, one would only borrow information from the green and blue clusters in the external data because they are shared with the trial data. However, one should avoid borrowing information from the purple cluster, as it is unique to the external data since borrowing from this cluster may introduce bias into the estimated treatment effect.

A brief review of the statistical model of SAM is provided next. For more detail refer to Bi and Ji (2023). Denote $Z_{i,j}$ as the cluster membership indicator for patient i in group j, where $\{Z_{i,j} = k\}$ indicates that patient i in group j is assigned to cluster k. The Bayesian nonparametrics model in SAM is given by a hierarchical structure as follows:

$$
\begin{aligned}
&x_{i,j}|Z_{i,j}, \{\phi_k\}_{k=1}^{\infty} \sim G\left(\phi_{Z_{i,j}}\right) \\
&Z_{i,j}|\{\pi_{j,k}\}_{k=1}^{\infty} \sim \sum_{k=1}^{\infty} \pi_{j,k}\delta_k(Z_{i,j}),\ \pi_{j,k} = \pi'_{j,k}\prod_{l=1}^{k-1}(1-\pi'_{j,l}) \\
&f(\pi'_{j,k}|\beta, \alpha_0, p_j) = p_j \times f_{\text{Beta}}\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right) + \underbrace{(1-p_j)\times I(\pi'_{j,k}=0)}_{(*)} \\
&p_j|a,b \sim Beta(a,b),\ \beta|\gamma \sim GEM(\gamma),\ \phi_k \sim H,
\end{aligned}
\tag{3.1}
$$

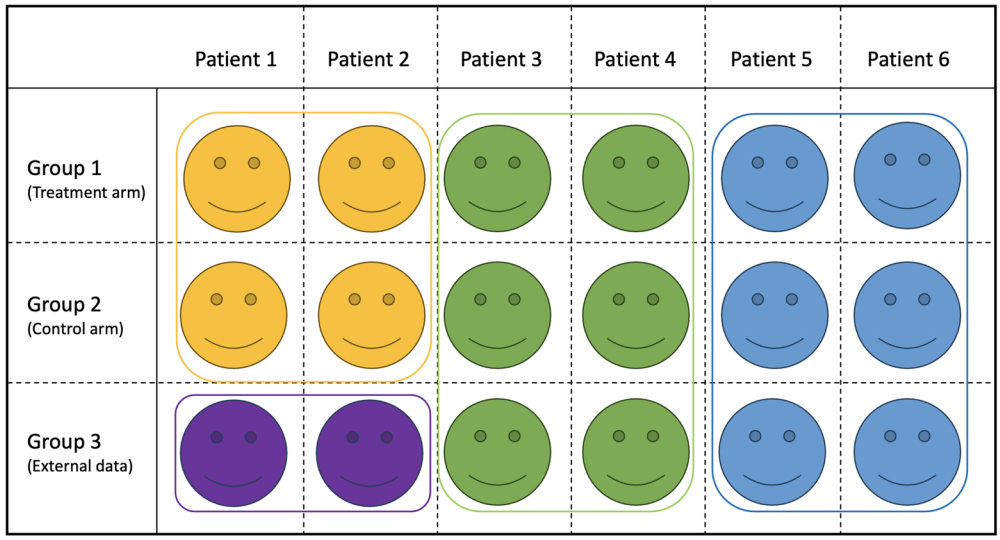

Figure 1 An illustration of clustering pattern under SAM. Rows represent groups and columns are patients within each group. The 3 groups correspond to the current RCT's treatment and control arms, and the external data. There are 4 homogeneous subpopoulations of patients (clusters) represented by colored smiley faces in blue, green, purple, and yellow. The boxes represent the common or unique clusters. For example, the green cluster is common and shared across all 3 groups, while purple is unique to group 3.

where G is some parametric distribution, $\delta_A(B)$ is the indicator function ($\delta_A(B) = 1$ if $B \in A$ or $B = A$, and $\delta_A(B) = 0$ otherwise), $f_{\text{Beta}}(a, b)$ is the density function of the Beta(a, b) distribution with mean $a/a + b$, $GEM(\gamma)$ represents the Griffths, Engen, and McCloskey distribution (Pitman 2002), and H is the base measure. For simplicity, we focus on continuous covariates, with $\phi_k = (\mu_k, \Sigma_k)$, $G(\phi_{Z_{i,j}}) = MVN\left(\mu_{Z_{i,j}}, \Sigma_{Z_{i,j}}\right)$ where MVN stands for the multivariate normal distribution, and $H = NIW(\mu_0, \lambda, \Psi, \nu)$ where NIW stands for to the normal-inverse-Wishart distribution. Parameter $\pi_{j,k}$ represents the cluster weights of cluster k in group j, and parameters (μ_k, Σ_k) denote (mean, covaraince matrix) of the kth cluster. Additional priors can be assigned to hyperparameters γ and α_0. For covariates of mixed types, such as multivariate data with continuous, binary and categorical variables, the model can be easily extended using a product likelihood for $G(\cdot)$ and for the base measure H. For details, see Chandra et al. (2023). Model (3.1) in SAM largely resembles the well known hierarchical Dirichlet Process (HDP) model (Teh et al. 2004), which induces common clusters across groups. SAM adds a unique model component $(*)$, which allows some common cluster to have zero weight in group j, thereby producing unique clusters.

Through model (3.1), SAM generates a joint posterior distribution of all the parameters including the cluster membership (ragged) matrix $Z = \{Z_{i,j}; i = 1, ., n_j, j = 1, ., J\}$, with which one could form the common and unique clusters. For example, a common cluster k between groups j and j' refers to the subpopulations of patients that share the same cluster label k. Since a posterior distribution of Z is generated, the number of clusters and clustering memberships themselves are random. In SAM-HC, we utilize the features of SAM that identifies common and unique clusters, upon which we build models and inference for constructing a hybrid control arm and estimating treatment effects.

4. SAM-HC

4.1. Clustering of patients

Let $A_{j,k} = \{i; Z_{i,j} = k, i = 1, \ldots, n_j\}$ represent the set of patients in the k-th cluster in group j. Also denote the set of cluster labels in each group j as $O_j = \{k; A_{j,k} \neq \emptyset\}$. We define the set of common cluster labels between a pair of groups as $C_{j,j'} = \{k; k \in O_j \cap O_{j'} \text{ for } j \neq j', j, j' \in \{1, 2, 3\}\}$. Our

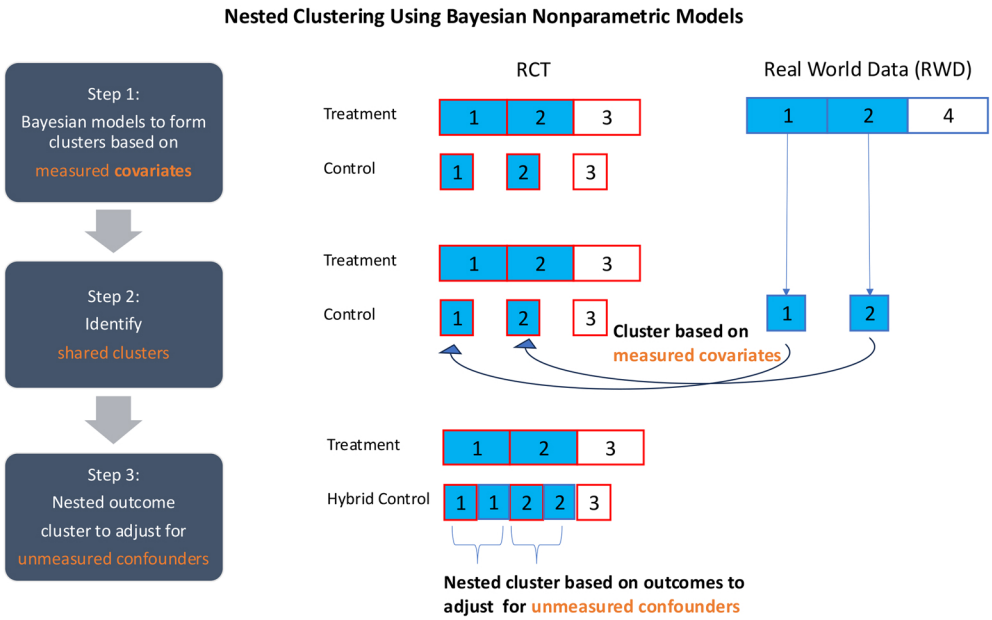

Figure 2 A stylized illustration of SAM-HC. Numbers in the boxes denote cluster labels. Boxes in red color represent patients in the RCT and those in blue represent patients in the external data. Cluster 4 is unique to the external data and therefore is not used for forming the HC. Cluster 3 is unique to the RCT and therefore is not augmented.

focus is on set $C_{2,3}$, which consists of the common cluster labels between the trial control arm and the external data. The proposed SAM-HC method augments $A_{2,k}$ by borrowing information from patients in $A_{3,k}$ for cluster(s) $k \in C_{2,3}$.

Figure 2 provides a schematic overview of SAM-HC. We consider a hypothetical example to illustrate the basic idea. Suppose a hypothetical cluster membership is given by

$$Z = \begin{bmatrix} 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 3 & & & \\ 1 & 1 & 2 & 2 & 4 & 4 \end{bmatrix},$$

where the rows are groups and columns are patients. Based on Z, there are a total of 4 clusters, with clusters $k = 1, 2, 3$ in groups 1 and 2, and clusters $k = 1, 2, 4$ in group 3. Clusters 1 and 2 are shared across groups 1 and 2, while cluster 4 is unique for group 3. Also, the set of patients in each cluster is given by $A_{1,1} = \{1, 2\}$, $A_{1,2} = \{3, 4\}$, $A_{1,3} = \{5, 6\}$, and $O_1 = \{1, 2, 3\}$ for group 1, $A_{2,1} = \{1\}$, $A_{2,2} = \{2\}$, $A_{2,3} = \{3\}$, and $O_2 = \{1, 2, 3\}$ for group 2, and $A_{3,1} = \{1, 2\}$, $A_{3,2} = \{3, 4\}$, $A_{3,4} = \{5, 6\}$, and $O_3 = \{1, 2, 4\}$ for group 3. The sets of common clusters are $C_{1,2} = \{1, 2, 3\}$, $C_{1,3} = \{1, 2\}$, and $C_{2,3} = \{1, 2\}$ between the treatment and control arms, between the treatment arm and the external data, and between the control arm and the external data, respectively. We focus on the set $C_{2,3} = \{1, 2\}$, and for clusters $k \in C_{2,3}$, ie $k = 1$ and $k = 2$, construct an HC by borrowing information from patients in the external data belonging to clusters 1 and 2, but not cluster 4. For illustrative purposes, the stylized example assumes all the clustering memberships are fixed. SAM-HC generates random clustering memberships to accommodate variabilities in statistical inference. This will be clear in Section 4.4 later.

In practice, the clustering distributions in the treatment and control arms of the RCT should be highly similar due to randomization. In rare cases where a common cluster is shared between the treatment arm of the RCT and external data, but not the control group, SAM-HC will not use the

common cluster to augment the control arm. For instance, if Z is given by

$$Z = \begin{bmatrix} 1 & 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & & & \\ 1 & 1 & 2 & 2 & 4 & 4 \end{bmatrix},$$

we would not borrow information from external data's cluster 4. One might allow information borrowing if there is a strong belief that cluster 4 is a meaningful subpopulation of the RCT.

4.2. Information borrowing across common clusters

4.2.1. Overview

We consider a sampling model for the observed outcomes and connect to the SAM model for covariates in model (3.1). Let Φ denote all the parameters in the SAM hierarchical model in (3.1) and write its posterior as $p(\Phi|x)$. Here Φ includes all the cluster labels, weights and other nuisance parameters in model (3.1). The outcome model is given by $f(y|\Theta, \Phi)$, where Θ represents the parameters in the outcome model including the key response parameters. We propose a prior distribution $\pi(\Theta|\Phi)$ for Θ including a power prior that allows information borrowing between the external data and control arm in the RCT. Then we perform inference on the response parameters based on the posterior $p(\Theta|y, x) \propto \int f(y|\Theta, \Phi)\pi(\Theta|\Phi)p(\Phi|x)d\Phi$. Note here we assume conditional on the clustering parameters Φ, the covariates x are independent of y, the outcome. Such an assumption is general in random effects models and latent variable Bayesian models. We incorporate the effect of outcome on borrowing through a power prior instead, which will be clear next. The proposed inference is conducted in 2 stages. First, obtain the posterior samples of $p(\Phi|x)$ based on the SAM in model (3.1), and then estimate the posterior of response parameters $p(\Theta|y, x)$.

4.2.2. Sampling model for outcome

For cluster k, denote $\theta_{1,k}$ and $\theta_{2,k}$ the cluster-specific response parameter in the treatment and control arms, respectively. Recall $y_1 = \{y_{i,1}; i = 1, \ldots, n_1\}$ are the observed patient responses in group 1, the treatment arm in the RCT. We assume

$$y_{i,1}|Z_{i,1} = k, \theta_{1,k} \sim f(y_{i,1}|\theta_{1,k}),\ \theta_{1,k}|\theta^* \sim \pi_0(\theta_{1,k}|\theta^*), \tag{4.2}$$

where $f(y_{i,1}|\theta_{1,k})$ denotes the p.d.f (sampling model) of y indexed by $\theta_{1,k}$ and $\pi_0(\theta_{1,k}|\theta^*)$ is the p.d.f of a vague prior for the parameter $\theta_{1,k}$ with hyperparameters θ^*. For continuous outcome, $F(\theta) = N(\mu, \sigma^2)$, $\theta_{1,k} = (\mu_{1,k}, \sigma^2_{1,k})$, and $\pi_0(\theta^*) = NIG(\mu_0, \nu_y, a_c, b_c)$ where NIG is the normal-inverse gamma distribution. For binary outcome, $F(\theta) = Bern(q)$, with $\theta_{1,k} = q_{1,k}$ and $\pi_0(\theta^*) = Beta(a_b, b_b)$. Recall $y_2 = \{y_{i,2}; i = 1, \ldots, n_2\}$ and $y_3 = \{y_{i,3}; i = 1, \ldots, n_3\}$ are the observed responses in groups 2 (the control arm of the RCT) and 3 (the external data), respectively. Assume

$$y_{i,2}|Z_{i,2} = k, \theta_{2,k} \sim f(y_{i,2}|\theta_{2,k}),$$

and the prior of $\theta_{2,k}$ is given by

$$p(\theta_{2,k}|y_3, A_{3,k}, \alpha_k, \theta^*) \propto \left[\prod_{i \in A_{3,k}} f(y_{i,3}|\theta_{2,k})\right]^{\alpha_k} \pi_0(\theta_{2,k}|\theta^*) \tag{4.3}$$

where $\alpha_k \in [0, 1]$ is a discount factor (or power parameter) for cluster k. The same $f(\cdot)$ and $\pi_0(\cdot)$ as the treatment arm are used here. Model (4.3) defines a power prior for $\theta_{2,k}$ that allows information within the same cluster k to be borrowed between groups 2 and 3.

4.2.3. Determine the power parameter α_k

Parameter α_k controls the degree of borrowing between the external data and control for cluster k. We determine α_k using a data-driven approach (similar in spirit to empirical Bayes). Specifically, we estimate α_k as a deterministic function of 1) the cluster weights $\pi_{j,k}$, 2) the cluster membership (ragged) matrix Z, and 3) the outcomes of the control and external data in cluster k. First, if $k \in O_3$

and $k \notin C_{2,3}$, or if $k \notin O_3$ and $k \in C_{2,3}$, we set $\alpha_k = 0$. In words, there is no borrowing across unique clusters in either the external data or the RCT data. Otherwise, the cluster is shared between the external data and control, which we would like to borrow information within this cluster. We restrict the amount of information borrowing to some maximum sample size $I = \frac{r-1}{r+1}N$, where N is the total sample size of both arms in the RCT and r is the ratio of the sample size of the treatment arm over the control arm. For example, if $N = 300$ and $r = 2$, then $I = 100$, which means one would borrow information of up to 100 patients from the external data to form an HC so that the amount of information in the HC matches that of information in the treatment arm.

Under this setting, the maximum value of the discount factor α_k in cluster k, denoted as $\alpha_{k,\max}$, is computed using the cluster membership and weights following Chen et al. (2020). Denote $\pi^*_{2,k}$ the desired proportion to be borrowed from the I patients for cluster k, and the maximum discount factor is computed as follows:

$$\alpha_{k,\max} = \min\left(\pi^*_{2,k} \cdot I,\ n_{3,k}\right)/n_{3,k}, \tag{4.4}$$

where $n_{3,k} = |A_{3,k}|$ is the number of elements in the set $A_{3,k}$ and $\pi^*_{2,k}$ is computed to match the sample size of the treatment arm and HC arm as follows. Recall $\pi_{j,k}$ is the probability (or weights) of cluster k in group j in the SAM model (3.1). We augment the control to form an HC in which the information is worth the same number of patients as in the treatment arm for each cluster k. Mathematically, this means

$$N \cdot \pi_{1,k} \cdot \frac{r}{r+1} = N \cdot \pi_{2,k} \cdot \frac{1}{r+1} + I \cdot \pi^*_{2,k}, \quad \text{where} I = \frac{r-1}{r+1}N.$$

Solving for $\pi^*_{2,k}$, we have

$$\pi^*_{2,k} = \frac{1}{r-1}(r\pi_{1,k} - \pi_{2,k}). \tag{4.5}$$

Equation (4.5) leads to a solution for (4.4), and hence a value for $\alpha_{k,\max}$. In practice, to prevent negative values of $\pi^*_{2,k}$ (when the control arm already has a larger number of patients in cluster k than the treatment arm), we use $\pi^*_{2,k+}$, ie $\pi^*_{2,k+} = \pi^*_{2,k}$ if $\pi^*_{2,k} > 0$, and $\pi^*_{2,k+} = 0$ otherwise.

Lastly, we connect α_k with outcome to account for potential unmeasured confounders. Specifically, within each cluster k, we compute the overlapping coefficient between the empirical distributions of the outcomes for the external data and the control arm of the RCT, and require the amount of information borrow, measured by α_k to be no larger than the overlapping coefficient. Let $y_{2,k} = \{y_{i,2}; Z_{i,2} = k\}$ and $y_{3,k} = \{y_{i,3}; Z_{i,3} = k\}$ denote the set of outcome measurements in cluster k for the control arm of the RCT and external data, respectively. Denote $f_{k,2}$ and $f_{k,3}$ their kernel density estimates. The overlapping coefficient (Inman and Bradley 1989) for cluster k is defined as:

$$\rho_k = \int_0^1 \min[f_{k,2}(y), f_{k,3}(y)]dy. \tag{4.6}$$

The discount factor α_k is then defined as:

$$\alpha_k = \min[\alpha_{k,\max}, \rho_k]. \tag{4.7}$$

The construction of $\alpha_{k,\max}$ in (4.4) and $\pi^*_{2,k}$ in (4.5) adaptively borrows information for cluster k based on the imbalance in the patient assignment between the treatment and control in cluster k. The construction of ρ_k further incorporates similarities of outcomes when covariates are believed to be matched (since these patients belong to the same cluster k). Therefore, (4.7) quantifies α_k as a function of similarities of both covariates and outcomes, partially accounting for bias caused by measured and unmeasured confounders.

4.3. Estimate treatment effects

Due to randomization, we assume $C_{1,2} = O_1 = O_2$, ie the treatment and control arms share all the clusters and there are no unique clusters in each of the 2 arms. Under this setting, denote Δ_k the cluster-specific treatment effect, for $k \in C_{1,2}$, given by

$$\Delta_k = \theta_{1,k} - \theta_{2,k}.$$

In rare cases where there are unique clusters in the treatment or control arms, we use an ad-hoc rule to merge the unique clusters to a common cluster in $C_{1,2}$ that has the shortest distance in terms of L2-norm between the cluster means. The overall treatment effect can be computed as a weighted average of the cluster-specific treatment effects Δ_k. Conditional on Z, we let

$$\Delta(Z) = \sum_{k \in O_1} \pi_{1,k} \Delta_k = \sum_{k \in O_1} \pi_{1,k}(\theta_{1,k} - \theta_{2,k}) \tag{4.8}$$

be the conditional overall treatment effects. We could either use $\{\pi_{1,k}\}$ or $\{\pi_{2,k}\}$ as the weights, which are in principle close to each other due to randomization. However, we decide to use $\{\pi_{1,k}\}$ since the treatment arm (group $j = 1$) is expected to have more patients and therefore lead to more stable estimates of clustering weights. In other words, our target estimand is the treatment effect standardized to the treated-arm covariate distribution. Because the hybrid control arm incorporates external observations, it can induce mixture shift relative to the treated arm. Averaging both group-specific mean outcomes over the same target (treated-arm) covariate distribution avoids this mismatch and yields an effect interpretable for the trial's treated population. In SAM, the weight $\pi_{1,k}$ (treated-arm cluster proportions) are estimated directly. As an alternative, these weights can be obtained follow Gelfand and Kottas (2002) or using a Bayesian bootstrap approach. The (unconditional) overall treatment effect is given by $\Delta = E[\Delta(Z)]$.

4.4. Inference

Bi and Ji (2023) develop a slice sampler to generate posterior samples via Markov chain Monte Carlo (MCMC) simulations. We use $m = 1, \ldots, M$ to index the M MCMC samples and use a generic notation $\hat{X}^{(m)}$ to denote the m-th sample for random variable X. Note that the posterior mean of overall treatment effect can be expressed as an integration of (4.8) over the posterior distributions of θ's, π, and Z, ie

$$\begin{aligned} \mathrm{E}[\Delta|y] &\equiv \mathrm{E}\left[\mathrm{E}[\Delta(Z,\pi)|Z,\pi,y]|x\right] \\ &= \int \left\{ \int \sum_{k \in O_1} \pi_{1,k}(\theta_{1,k} - \theta_{2,k}) p(\theta_{1,k}|y,Z) p(\theta_{2,k}|y,Z) d\theta \right\} p(\pi_1, Z|x) d\pi_1 dZ \end{aligned} \tag{4.9}$$

where $\theta = \{(\theta_{1,k}, \theta_{2,k}) : k \in O_1\}$, and $\pi_1 = \{\pi_{1,k} : k \in O_1\}$. As mentioned in Section 4.2, (4.9) assumes that conditional on x, π and Z are independent of y. Using the MCMC samples, the posterior mean (4.9) is estimated as follows. For the m-th sample, let $\hat{O}_j^{(m)}$ be the cluster labels for group j. Let cluster $k^{(m)} \in \hat{O}_1^{(m)}$, then we compute the m-th posterior sample of the cluster-specific treatment effect as

$$\hat{\Delta}^{(m)}_{k^{(m)}} = \hat{\theta}^{(m)}_{1,k^{(m)}} - \hat{\theta}^{(m)}_{2,k^{(m)}}, \tag{4.10}$$

where $\hat{\theta}^{(m)}_{j,k^{(m)}}$ for $j = 1, 2$ are 2 random samples from the posterior distributions of $\theta_{j,k}$ obtained from models in Section 3.2, conditional on the cluster membership $Z^{(m)}$ sampled in the mth MCMC iteration. Finally, the overall treatment effect $\hat{\Delta}^{(m)}$ can be obtained with $\hat{\Delta}^{(m)}_{k^{(m)}}$ and the weights $\hat{\pi}^{(m)}_1$ using (4.8):

$$\hat{\Delta}^{(m)} = \sum_{k^{(m)} \in \hat{O}_1^{(m)}} \hat{\pi}^{(m)}_{1,k^{(m)}} \hat{\Delta}^{(m)}_{k^{(m)}}, \tag{4.11}$$

Algorithm 1 SAM-HC

for $m = 1, \cdots, M$ **do**
  Estimate $\boldsymbol{Z}^{(m)}$ and $\boldsymbol{\pi}_1^{(m)}$ conditional on $\boldsymbol{x}$ using the slice sampler proposed in (2023).
  Compute $A_{j,k}$ and O_j for $j = 1, 2, 3$ and $k = 1, \ldots, |O_j|$ from $\boldsymbol{Z}^{(m)}$. Also compute $C_{2,3}$.
  for each $k \in C_{2,3}$ **do**
    if $k \notin O_3$ **then**
      Set $\alpha_k = 0$.
    else
      Compute ρ_k with equation (4.6) and $\alpha_{k,\max}$ using equation (4.4). Set $\alpha_k = \min[\alpha_k, \rho_k]$.
    end if
    Draw random samples $\hat{\theta}^{(m)}_{1,k^{(m)}}$ and $\hat{\theta}_{2,k^{(m)}}$ from the posterior distributions of $\theta_{1,k}$ (model (4.2)) and $\theta_{2,k}$ (model (4.3)), and compute $\hat{\Delta}^{(m)}_{k^{(m)}}$ using equation (4.10).
  end for
  Compute $\hat{\Delta}^{(m)}$ using equation (4.11).
end for

and the posterior mean treatment effect is estimated as $\sum_{m=1}^{M} \hat{\Delta}^{(m)}/M$. Also, given the posterior sample $\{\hat{\Delta}^{(m)};\ m = 1, \ldots, M\}$, we can easily compute various quantities of interest, such as the standard deviation of the overall treatment effect. Additionally, it allows us to determine the posterior probability of a significant treatment effect, denoted as

$$\Pr(\Delta > \epsilon | \text{Data}),$$

for some minimal clinically meaningful treatment effect ϵ.

Algorithm 1 shows a summary of the proposed SAM-HC method. A detailed MCMC pseudo-code is provided in Appendix Section A.2, supplementary material.

5. Simulation studies

5.1. Simulation setup

We generate 3-dimensional baseline covariates for the current RCT (groups $j = 1$ and $j = 2$) and for the external data (group $j = 3$) from a 4-component mixture of multivariate normal distributions under 3 scenarios. Specifically, we assume

$$f(x) = \sum_{k=1}^{4} w_k \times MVN(x | \mu_k, \Sigma_k),$$

where MVN denotes a multivariate normal density. Throughout, $\Sigma_k = I$ for all k, and $\mu_k \in \mathcal{R}^3$. The component means are $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{2_3, 0_3, -2_3, -4_3\}$ (Notation: $2_3, 0_3, -2_3$, and -4_3 denote 3-dimensional vectors with all entries equal to 2, 0, –2, and –4, respectively).

For the current RCT covariates, the mixture is fixed across scenarios with weights $\{w_1, w_2, w_3, w_4\} = \{0.3, 0.4, 0.3, 0.0\}$. Thus, the current RCT contains 3 clusters with means at 2_3, 0_3, and -2_3.

For the external data, we vary the weights by scenario to induce different overlap patterns with the RCT:

- Scenario 1 (Superset): $(w_1, w_2, w_3, w_4) = (0.2, 0.3, 0.3, 0.2)$. The external data includes an *additional cluster* at -4_3 and allocates different weights to the clusters common with the RCT.
- Scenario 2 (Overlap): $(w_1, w_2, w_3, w_4) = (0.5, 0.3, 0.0, 0.2)$. The external data shares the clusters at 2_3 and 0_3 with the current RCT and also includes a unique cluster at -4_3.

- Scenario 3 (Subset): $(w_1, w_2, w_3, w_4) = (0.5, 0.5, 0.0, 0.0)$. The external data is a subset of the RCT distribution, sharing only the clusters at 2_3 and 0_3, with no unique clusters.

Outcomes y are associated with the baseline covariates. We consider both continuous and binary outcomes.

For continuous outcomes, we simulate the outcome $y_{i,j}$ using a linear regression given by:

$$y_{i,j} \sim \beta_{0,j} + \beta^T x_{i,j} + \epsilon_{i,j},\ \epsilon_{i,j} \sim N(0,1),\ j = 1, 2, 3.$$

For binary outcomes, we generate $P(y_{i,j} = 1|x_{i,j})$ using a logistic regression given by:

$$\text{logit}\, P(y_{i,j} = 1|x_{i,j}) = \beta_{0,j} + \beta^T x_{i,j},$$

and then simulate $y_{i,j}$ from the Bernoulli distribution with the success probability $P(y_{i,j} = 1|x_{i,j})$.

In the simulations, the intercept varies by arm (treatment, control, external) via $\beta_{0,j}$ where $j \in \{1, 2, 3\}$ indexes the arm, while the regression slopes β are shared across arms. Specifically, we fix $\beta_{0,2} = \beta_{0,3} = 0$, and set $\beta = 1_3$. We vary $\beta_{0,1} \in \{0, 1, 2, 3\}$ to induce different treatment effects. Note that neither the regression slopes nor the intercepts vary across clusters. Furthermore, we assume that the current RCT uses a randomization ratio of $r = 2:1$. We simulate the RCT data with 2 different sample sizes, $N = \{300, 450\}$. For the case where $N = 300$, we place 200 patients in the treatment arm and 100 in the control. We generate a total of 300 patients for the external data. For the case where $N = 450$, we place 300, 150, and 450 patients in the treatment arm, control arm, and the external data, respectively.

The parameter of interest is the treatment effect Δ, which represents the difference in mean responses between the treatment and control arms. Consequently to $\beta_{0,1} = 0$, 1, 2, and 3, the true values of Δ are 0, 1, 2, and 3, respectively, for the continuous outcome, and 0.00%, 9.42%, 15.76%, and 19.52%, respectively, for the binary outcome.

We compare the proposed SAM-HC method with 3 other methods: a baseline method without borrowing and 2 methods that borrow information from external data using the baseline covariates. Specifically, the baseline method estimates the treatment effect using only the RCT data and does not borrow information from the external data. Inference of treatment effect is based on the maximum likelihood estimation (MLE). The second method is the PSCL method proposed by Chen et al. (2020). We implement the PSCL method using 3 different strategies, corresponding to the ways propensity scores are estimated. Namely, they are PSCL 1, PSCL 2, and PSCL 3, referring to the PSCL method with propensity scores estimated using a first-order logistic regression, random forest, and a logistic regression with model selection, respectively. All 3 versions of PSCL are implemented in the R package "PSRWE" (Wang and Chen 2022). Lastly, the third method is PS-MAP Liu et al. (2021), which models PS using the MAP prior. Finally, for each scenario, we simulate 1,000 datasets.

We use the following priors and hyperparameters for the proposed SAM-HC method. For SAM model (3.1), we set $a = b = 0.5$, and utilize a $Gamma(3, 3)$ prior for the hyperparameters γ and α_0. Additionally, we set $\mu_0 = 0_3$, $\Psi = I$, $\lambda = 0.1$, and $\nu = 3$. These hyperparameters are standard settings to construct vague priors for normal and inverse Wishart distributions. Note here the value 3 is the number of covariates in the simulation. For datasets with p covariates, we recommend setting $\mu_0 = 0_p$, $\Psi = I_{p\times p}$, and $\nu = p$. Additionally, the choice of $a = b = 0.5$ leads to a Jeffrey's prior $Beta(0.5, 0.5)$ which is considered a default choice of a vague prior. The use of $Gamma(3, 3)$ has been discussed in Bi and Ji (2023), again as a choice of vague prior. For the power priors in SAM-HC, we adopt a normal distribution $N(\mu, \sigma^2)$ for the continuous outcome and a Bernoulli distribution Bern(p) for the binary outcome. Conjugate priors are chosen for the parameters in the sampling model. Specifically, we use a normal-inverse-gamma prior $NIG(0, 0.1, 3, 3)$ for (μ, σ^2) and $Beta(0.5, 0.5)$ for p. These are standard priors that are not informative and routinely applied in the literature. Lastly, we run an MCMC simulation of 10,000 iterations, with burn-in period of 5,000 iterations.

We summarize the posterior cluster membership using an optimal clustering method (Meilă 2007) to obtain a point estimate. To assess the clustering accuracy in comparison to the ground

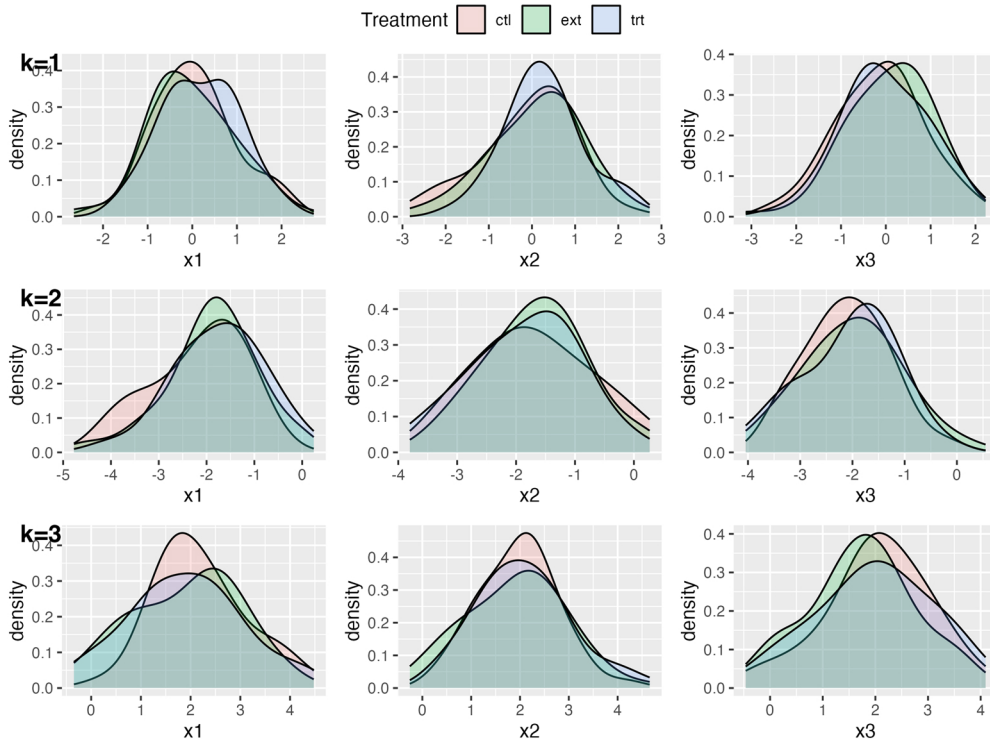

Figure 3 The covariate density plots of one simulated data in Scenario 1. The rows represent 3 clusters estimated by SAM-HC.

truth cluster membership of each patient, we use the adjusted Rand index (ARI) (Hubert and Arabie 1985) and the normalized Frobenius distance (NFD) (Horn and Johnson 1990). More detail can be found in Bi and Ji (2023). Lastly, we assess the performance of SAM-HC in terms of the estimated overall treatment effect, including the mean, standard deviation, bias, and mean squared error (MSE) across all simulated datasets. In Appendix Section A.4, supplementary material, we also assess SAM-HC's covarage probability of its 95% credible interval, type I error rate, and statistical power to better understand the operating characteristics of the proposed method.

5.2. Simulation results

We assess similarity in the distributions of covariates between the treatment and HC. We first check that under SAM-HC, if the distribution of covariates of the treatment arm is similar to the distribution of the hybrid control arm. Specifically, within each estimated cluster, the distributions of covariates should be similar between the 2 arms. We randomly selected 1 dataset in each scenario with $N = 300$. We plot the density of the covariates by estimated clusters in Fig. 3 for Scenario 1, and in Figs. 1 and 2 in Appendix, supplementary material for scenarios 2 and 3, respectively.

Figure 3 shows that the distribution of each covariate are indeed similar between the treatment arm, control arm, and the external data, for each inferred cluster k. Furthermore, the estimated cluster centers are shown in Appendix Table 3, supplementary material. In addition, the cluster-specific treatment effects for the 3 selected datasets are reported in Tables 4 and 5 in Appendix, supplementary material. In the selected examples, we see that SAM-HC is able to correctly identify the number of clusters in these selected examples. The estimated cluster centers are also close to the true values in their corresponding scenarios. In addition, Tables ?? and ?? suggest that treatment effects for clusters are well estimated.

To further assess the similarity of covariate distributions between treatment and the hybrid control arms, we follow the procedure outlined in Chandra et al. (2023) and apply the Bayesian

Additive Regression Tree (BART) model (Chipman et al. 2010). For each estimated cluster k, we aggregate data from all 3 groups and create a dummy variable T_i indicating whether patient i belongs to the treatment arm ($T_i = 1$) or the hybrid control arm ($T_i = 0$). We then carry out a 10-fold cross-validation, with 9-folds used as the training data and 1-fold as the testing data. We apply BART to predict whether an observation in the testing data belongs to the treatment arm or not. The results are reported as the Area Under the ROC Curve (AUC). A value around 0.5 indicates no difference between the patients in the current treatment arm and the patients in the hybrid control. We randomly select 10 datasets in each scenario, and the corresponding results show that across all scenarios, the range of mean AUC values is between 0.525 and 0.544, all around 0.5. This indicates that the covariate distributions for the treatment and hybrid control arms are similar and indistinguishable by BART.

Next, we report the clustering results of SAM-HC for all simulated datasets. The true number of clusters is 4 in Scenario 1 and Scenario 2, and 3 in Scenario 3. For ARI and NFD, the closer the value of ARI is to 1 or the value of NFD to 0, the better the clustering result of the method. Table 6 in Appendix, supplementary material, shows the estimated total number of clusters across all groups, as well as the ARI and the NFD of the estimated clusters compared to the true cluster membership. On average, the number of estimated clusters is accurate, close to its truth in all cases. The ARI and NFD values are satisfactory, improving with increasing sample size.

We present the estimated treatment effect, its standard deviation, and the mean squared error (MSE). We compare these results with the baseline method, PSCL 1-3, and PS-MAP. Table 2 and Table 7 in Appendix, supplementary material, provide a summary of the results for continuous and binary outcomes, respectively. Tables 8 and 9 in Appendix, supplementary material, provides the coverage probabilities, type I error rates, and powers across the scenarios for these 2 types of outcomes.

5.2.1. Scenario-specific findings

5.2.1.1. Scenario 1 (superset).

In this setting, although SAM-HC achieves smaller MSE than the baseline and the PS-MAP methods, PSCL 1 attains lower MSE than SAM-HC, and PSCL 3 is performed on par with SAM-HC. Given that PSCL methods run substantially faster, SAM-HC does not offer a clear efficiency advantage here. We therefore view SAM-HC as competitive—but not preferred—in this scenario when judged by MSE and computation time.

5.2.1.2. Scenario 2 (overlap).

When the external data share some, but not all, clusters with the current RCT, SAM-HC performs best overall in our simulations (lowest MSE at $N = 300$ with further improvement at $N = 450$, see Table 2 and Appendix Table 7, supplementary material). This aligns with the model's design goal: to borrow from overlapping subpopulations while excluding non-overlapping ones. To illustrate adaptivity, we report the inclusion probability—the posterior probability that an external-data patient is borrowed into the hybrid control—which we define as $\Pr(Z_{i,3} \in C_{2,3} \mid D)$. For the unique external cluster in Scenario 2, mean inclusion probabilities are low (eg 3.47%, SD = 0.17 for $N = 300$ and 2.51%, SD = 0.15 for $N = 450$, respectively), whereas for shared clusters they exceed 91%, indicating selective borrowing consistent with the intended construction.

5.2.1.3. Scenario 3 (subset).

Although SAM-HC shows the lowest bias, its MSE is larger than that of PSCL 1, with the gap narrowing at larger N. In settings with strong covariate overlap and when computation time is a priority, PS-based comparators therefore offer a better bias-variance trade-off, while SAM-HC remains a competitive option when selective, cluster-aware borrowing is desired.

We also evaluate the performance of SAM-HC with binary outcomes, and the results are shown in Table 7 in Appendix, supplementary material. Similar to the continuous outcome, SAM-HC achieves the best performance in Scenario 2, and can be a competitive method in other scenarios.

Table 2 Simulation results based on continuous outcome for SAM-HC, baseline method, 3 versions of PSCL methods, and PS-MAP method. Here $\hat{\Delta}$ is the average posterior mean of the overall treatment effect across 100 simulated trials.

N	Sc	Method	True $\Delta = 0$			True $\Delta = 1$			True $\Delta = 2$			True $\Delta = 3$		
			$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE
300	Sc 1	Baseline	0.02	0.61	0.37	1.02	0.61	0.37	2.02	0.61	0.37	3.01	0.61	0.37
		PSCL1	0.03	0.18	0.03	1.03	0.19	0.04	2.03	0.18	0.04	3.03	0.18	0.03
		PSCL2	0.52	0.46	0.48	1.52	0.46	0.49	2.53	0.46	0.49	3.52	0.46	0.48
		PSCL3	0.14	0.41	0.17	1.14	0.41	0.18	2.14	0.41	0.17	3.13	0.40	0.17
		PS-MAP	0.08	0.80	0.58	0.94	0.88	0.58	1.86	1.05	0.60	2.70	1.26	0.57
		SAM-HC	0.09	0.35	0.14	1.09	0.35	0.14	2.09	0.34	0.14	3.08	0.34	0.13
	Sc 2	Baseline	−0.01	0.59	0.35	1.00	0.59	0.35	2.00	0.59	0.35	3.00	0.60	0.36
		PSCL1	−0.02	0.34	0.11	0.97	0.33	0.11	1.98	0.33	0.11	2.97	0.34	0.11
		PSCL2	−0.48	0.35	0.35	0.51	0.35	0.36	1.52	0.34	0.35	2.52	0.35	0.35
		PSCL3	−0.39	0.34	0.27	0.61	0.34	0.27	1.62	0.34	0.26	2.61	0.34	0.26
		PS-MAP	−0.02	0.81	0.59	0.89	0.86	0.61	1.81	0.99	0.56	2.75	1.19	0.56
		SAM-HC	−0.02	0.23	0.05	0.98	0.23	0.05	1.98	0.23	0.05	2.98	0.23	0.05
	Sc 3	Baseline	0.03	0.60	0.37	1.03	0.60	0.36	2.04	0.60	0.36	3.03	0.60	0.37
		PSCL1	−0.07	0.17	0.03	0.93	0.18	0.04	1.93	0.17	0.03	2.93	0.17	0.03
		PSCL2	−0.45	0.32	0.30	0.56	0.31	0.29	1.56	0.31	0.29	2.55	0.31	0.29
		PSCL3	−0.16	0.25	0.08	0.84	0.24	0.08	1.85	0.25	0.08	2.84	0.25	0.08
		PS-MAP	−0.52	0.90	0.98	0.33	0.88	1.01	1.21	0.98	1.03	2.14	1.15	1.01
		SAM-HC	−0.00	0.24	0.07	0.99	0.24	0.06	2.00	0.24	0.06	2.99	0.23	0.06
450	Sc 1	Baseline	−0.00	0.49	0.24	1.00	0.49	0.24	2.00	0.49	0.24	3.00	0.49	0.24
		PSCL1	0.04	0.14	0.02	1.04	0.15	0.02	2.04	0.14	0.02	3.04	0.14	0.02
		PSCL2	0.52	0.38	0.41	1.52	0.38	0.42	2.52	0.38	0.41	3.52	0.38	0.42
		PSCL3	0.15	0.27	0.10	1.15	0.28	0.10	2.14	0.27	0.09	3.15	0.28	0.10
		PS-MAP	0.11	0.66	0.40	0.98	0.75	0.40	1.89	0.95	0.40	2.78	1.21	0.43
		SAM-HC	0.03	0.11	0.03	1.02	0.11	0.02	2.02	0.11	0.03	3.02	0.11	0.02
	Sc 2	Baseline	−0.01	0.49	0.24	0.99	0.49	0.24	1.99	0.49	0.24	2.99	0.49	0.24
		PSCL1	−0.02	0.26	0.07	0.98	0.26	0.07	1.98	0.26	0.07	2.98	0.27	0.07
		PSCL2	−0.46	0.28	0.29	0.54	0.28	0.30	1.54	0.29	0.30	2.54	0.29	0.30
		PSCL3	−0.39	0.28	0.23	0.61	0.29	0.23	1.61	0.29	0.23	2.61	0.28	0.23
		PS-MAP	0.01	0.70	0.43	0.91	0.76	0.45	1.83	0.93	0.43	2.69	1.17	0.44
		SAM-HC	0.00	0.17	0.03	1.00	0.17	0.03	2.00	0.17	0.03	2.99	0.17	0.03
	Sc 3	Baseline	0.02	0.51	0.27	1.01	0.52	0.27	2.01	0.52	0.27	3.01	0.52	0.27
		PSCL1	−0.08	0.13	0.02	0.92	0.14	0.02	1.92	0.14	0.03	2.92	0.14	0.03
		PSCL2	−0.44	0.25	0.24	0.56	0.26	0.24	1.56	0.26	0.24	2.56	0.26	0.24
		PSCL3	−0.17	0.18	0.06	0.83	0.18	0.06	1.83	0.19	0.06	2.83	0.19	0.06
		PS-MAP	−0.60	0.73	0.82	0.34	0.73	0.78	1.21	0.81	0.81	2.09	0.98	0.85
		SAM-HC	0.00	0.20	0.04	1.00	0.20	0.04	2.00	0.20	0.04	2.99	0.20	0.04

The proposed method incurs some computational burden. For example, in Scenario 1, SAM-HC takes around 300 seconds to analyze 1 dataset, compared to less than 60 seconds for the other methods.

Finally, in a sensitivity analysis, we tested the case where the assumptions of no unmeasured confounders and identical outcome models between the current control and external data do not hold. Specifically, the outcome model for the external data depends on an unmeasured confounder, resulting in differences between the outcome models of the external data and the control arm of the current RCT. However, we allow this unmeasured confounder to potentially correlate with an observed covariate, as is often the case in multivariate analysis (Chandra et al. 2023). Details are

provided in Appendix Section A.5, supplementary material, where SAM-HC is compared to the baseline method as well as a simple variant using the test-then-pool approach. The results indicate that both SAM-HC and its simple variant with the test-then-pool approach achieve similar biases as the baseline method, with an improved variance. This suggests some robustness of the borrowing strategy when modest outcome-model discrepancies are present.

Overall, SAM-HC's advantage is most evident when the external data partially overlap with the RCT in subpopulations (Scenario 2). In settings with broader alignment or simple subset relations (Scenarios 1 and 3), PS-based methods (PSCL 1-3) are competitive or superior in MSE and are more computationally efficient. We view SAM-HC as a complementary option: it provides transparent, cluster-level borrowing decisions and uncertainty quantification, with the trade-off of higher computational complexity.

6. Application

6.1. Application motivation

In randomized trials for conditions like atopic dermatitis, enrolling patients to receive placebo control becomes increasingly challenging as effective treatments become available. Sponsors often prefer $r:1$ randomization (with $r > 1$) to limit control enrollment while maintaining statistical power. SAM-HC enables responsible incorporation of high-quality external control data under harmonized eligibility criteria.

The workflow proceeds as follows: (i) apply SAM to identify common and unique patient subpopulations between trial and external data based on baseline covariates, (ii) borrow information only from matched subpopulations in external data, with borrowing capped to achieve effective 1:1 randomization, and (iii) estimate treatment effects using the hybrid control. This approach maintains trial integrity while enhancing precision when external and trial populations overlap.

6.2. Background and dataset

We consider clinical trials for patients with AD, a chronic inflammatory skin disease associated with substantial patient burden, characterized by recurrent eczematous lesions and intense pruritus (Simpson et al. 2022).

We analyze placebo (control) arms from 3 historical AD trials (NCT numbers NCT03569293, NCT03607422, and NCT03568318 for Trials 1, 2, and 3, respectively) with control-arm sizes 263, 265, and 306, respectively. Baseline characteristics include gender, age, race, ethnicity, body mass index (BMI), baseline body surface area affected (BSA), and baseline Eczema Area and Severity Index (EASI). Individual-level Week-16 EASI measurements were not available under data-use agreements; only the binary EASI-75 responder indicator was harmonized across trials, defined as:

- $\text{EASI}_{\text{pc}} = \frac{\text{Baseline EASI}-\text{16-week EASI}}{\text{Baseline EASI}} \times 100\%$;
- $y = 1$ if $\text{EASI}_{pc} \geq 75\%$, and 0 otherwise.

To emulate a realistic $2:1$ randomized trial while preserving real covariate-outcome structure, we create a semi-synthetic spike-in dataset: we designate the control arm of Trial 1 as the "treated" cohort (size 263) and sample without replacement 131 patients from Trial 2 to serve as the RCT control (total $N = 394$). Baseline covariates show no material imbalances between these 2 arms (Appendix Tables 11 and 12, supplementary material). Trial 3 (306 patients) supplies external placebo data used by SAM-HC to construct the hybrid-control via baseline-only clustering (to avoid outcome leakage). Observed EASI-75 response rates are of 25%, 21%, and 34%, for Trials 1 to 3, respectively; the across-trial mean is $\approx 28\%$ (null scenario).

To also study a non-null setting while preserving observed covariate-outcome relationships, we "spike in" a higher response in the treated cohort by an intercept shift on the logit scale. Concretely, let $\{(y_{i,1}, x_{i,1}); i = 1, \ldots, 263\}$ denote outcomes and covariates from Trial 1. Fit

$$\text{logit}\ \Pr(y_{i,1} = 1 | x_{i,1}) = \hat{\beta}_0 + \hat{\beta}^T x_{i,1}.$$

Table 3 Estimated cluster mean and cluster-specific treatment effect using the data of the AD trial.

		Cluster							
		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Weights	$\pi_{1,k}$	0.14		0.20		0.42		0.24	
	$\pi_{2,k}$	0.18		0.18		0.40		0.24	
	$\pi_{3,k}$	0.16		0.23		0.32		0.29	
Cluster mean		Mean	SD	Mean	SD	Mean	SD	Mean	SD
	Age	19.86	0.70	37.79	1.37	41.26	1.54	21.56	0.88
	Baseline EASI	18.55	0.46	39.04	1.53	21.68	0.40	33.82	1.87
	BSA	24.56	1.07	66.30	2.52	31.79	1.26	57.06	2.85
	BMI	22.78	0.50	29.11	0.79	27.30	0.74	22.12	0.40
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cluster-specific treatment effect	Null	0.088	0.121	–0.035	0.074	–0.085	0.071	–0.087	0.093
	Alternative	0.659	0.102	0.561	0.076	0.501	0.065	0.519	0.091

Table 4 Estimated treatment effects for the AD trial, using the proposed SAM-HC method, the baseline method, the 3 versions of PSCL method, and PS-MAP method.

	Method	Δ	SD	Significance
Null Case	Baseline	0.026	0.044	P-value: 0.281
	PSCL1	−0.045	0.149	P-value: 0.618
	PSCL2	−0.034	0.149	P-value: 0.590
	PSCL3	−0.041	0.315	P-value: 0.551
	PS-MAP	−0.040	0.038	$\Pr(\Delta > 0 \mid \text{data}) = 0.14$
	SAM-HC	−0.051	0.038	$\Pr(\Delta > 0 \mid \text{data}) = 0.09$
Alternative Case	Baseline	0.627	0.042	P-value: 0.000
	PSCL1	0.596	0.149	P-value: 8.4e-6
	PSCL2	0.606	0.139	P-value: 6.5e-6
	PSCL3	0.598	0.140	P-value: 1.0e-5
	PS-MAP	0.555	0.035	$\Pr(\Delta > 0 \mid \text{data}) = 1.00$
	SAM-HC	0.539	0.036	$\Pr(\Delta > 0 \mid \text{data}) = 1.00$

Define a shifted linear predictor $\hat{\beta}_0 + \delta + \hat{\beta}^T x_{i,1}$ and generate

$$\tilde{y}_{i,1} \sim Bern(p_i), \ \text{logit}\, p_i = \hat{\beta}_0 + \delta + \hat{\beta}^T x_{i,1}.$$

We choose the scalar intercept shift $\delta = 2.33$ (via a small grid search) so that $\frac{1}{263}\sum_i p_i \approx 0.80$. This yields an alternative (non-null) scenario with a ground-truth treated response of $\approx 80\%$ and an implied average treatment effect of roughly $80\% - 28\% \approx 52\%$ (alternative scenario).

6.3. Analysis results

Our primary estimand is the average treatment effect standardized to the treated-arm covariate distribution (see Section 3.3). The posterior mean number of clusters is 4.15 (SD = 0.36). For interpretation, we report a point estimate of the partition; this comprises 4 clusters that are common across all 3 arms, with no trail-unique clusters under the posterior mode of the co-clustering matrix. Table 3 summarizes cluster means and cluster-specific treatment effects (Δ_k) under the null and alternative scenarios. Table 4 compares overall treatment effects across methods (baseline no-borrowing; SAM-HC; PSCL 1-3; PS-MAP).

6.3.1. Null scenario

All methods yield effects close to zero, as expected. Methods that borrow (SAM-HC, PSCL, PS-MAP) yield slightly negative estimates because the external control response (34%) exceeds both the

emulated treatment (25%) and in-trial control (21%), which raises the hybrid-control rate and attenuates the treatment-control difference relative to the no-borrowing baseline. For SAM-HC, the posterior mean ATE is -0.051, with 95% CrI [-0.12, 0.02] and $\Pr(ATE > 0) = 0.09$.

6.3.2. Alternative scenario

All methods detect a positive effect. Because borrowing pulls the control rate towards 34%, SAM-HC, PSCL, and PS-MAP report smaller (but still positive) effects than the no-borrowing baseline. For SAM-HC, the posteior mean is 0.54 with 95% CrI [0.47, 0.61] and $\Pr(ATE > 0) = 1.00$. The spike-in calibratrion targets an $\approx$ 80% treated response; under our standardization the ground-truth ATE is $\approx$ 52% by construction (80% minus the standardized control mean), and the SAM-HC estimate is close to this target.

6.4. Practical implications and limitations

When assigning patients to placebo is less desirable—but an internal randomized comparison remain necessary—SAM-HC provides a principled hybrid-control that (i) borrows only from covariate-matched subpopulations, (ii) limits the total information borrowed via an ESS cap, and (iii) moderates borrowing when external outcomes are systematically higher than in-trial control. In settings like ours—where the external placebo response is higher—these safeguards temper over-optimistic treatment effects while retaining precision gains.

This application uses a derived binary endpoint (EASI-75) because individual-level continuous EASI at Week 16 was no accessible under the data-use agreements; results therefore target the EASI-75 estimand. The semi-synthetic 2 : 1 analogue retains real-world covariate structure but excludes sponsor-arm outcomes for confidentiality. We mitigate potential residual confounding by restricting to common clusters and moderating borrowing via outcome overlap; in practice, sponsors could add negative-control outcome checks or adopt a righter ESS cap when exchangeability is uncertain.

7. Discussion

In this study, we introduce the SAM-HC method to augment the control arm of an RCT using external data and improve the estimation of treatment effects. A key innovation is to identify common subpopulations of patients between the RCT and the external data and allow information to be borrowed only across these common subpopulations. Therefore, pooling information from these "similar" patients in the external data may help mitigate the risk of increased bias that can arise from pooling data from heterogeneous patients in the external data. The overlapping coefficient may help further down-weighting the external data to reduce potential bias caused by unmeasured confounders. Overall, SAM-HC is comparable to PS-based alternatives, and it performs particularly well when not all patient subpopulations are shared between RCT and external data. Thanks to the model-based inference on all the unknown parameters using BNP models, SAM-HC is powerful in reporting posterior distributions of cluster-specific treatment effects, overall treatment effects, and the random clusters themselves.

However, it is important to acknowledge the limitations of the current method. Firstly, the proposed method incurs substantial computation burden. SAM-HC's primary computational cost arises from MCMC sampling of the partition structure. For $n = 300$ patients, full Bayesian analysis requires approximately 2 to 3 hours for 10,000 iterations on standard hardware, compared to 5 to 10 minutes for PS-based methods. This trade-off provides full posterior uncertainty quantification, including clustering uncertainty. We recommend SAM-HC when: (i) external data contain distinct subpopulations not present in the trial, (ii) selective cluster-aware borrowing is critical, and (iii) full uncertainty quantification is required. PS-based methods may be preferred when: (i) populations strongly overlap, (ii) rapid analysis is needed, or (iii) all covariates are well-measured and confounding is minimal. Moreover, the assumption that covariates are continuous variables can be mitigated by a model extension. See Appendix Section A.6, supplementary material, for an illustration. Lastly, additional fine tuning might be desirable to accommodate potential unmeasured confounders. The proposed method only downweights the power parameter in

information borrowing based on the overlapping coefficients, rather than precisely limit borrowing to those patients with similar outcomes.

Finally, subjects belonging to different clusters may still exhibit similar outcomes. This could mean that the covariates used for clustering are not associated with outcome, or the errors associated with covariates are negligible compared to the errors associated with outcome. In the former case, one could perform a variable selection first to select relevant covariates. In the latter case, more sophisticated models might be needed. For example, one may consider a model-averaging approach that allows borrowing either when covariates are homogeneous or the predicted outcomes are homogeneous.(Schwartz et al. 2024).

8. Software

The R code used in this paper can be found on the GitHub repository https://github.com/edwardbi/SAMHC/tree/main. The repository includes R implementations of the MCMC sampler, example datasets, and tutorials for reproducing the simulation studies and application results.

9. Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org. This includes: (1) complete notation table, (2) MCMC algorithm pseudocode, (3) additional simulation scenarios with mixed-type covariates, (4) extended application results, and (5) computational timing comparisons.

Supplementary material

Supplementary material is available at *Biostatistics Journal* online.

Funding

None declared.

Conflict of interest

None declared.

Disclaimer

The external control data was based in part on data from the TransCelerate BioPharma Inc. Historical Trial Data (HTD) Sharing Initiative, which includes contributions of anonymized or pseudonymized data from TransCelerate HTD member companies including AbbVie, Amgen, Astellas, AstraZeneca, Boehringer Ingelheim, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Johnson & Johnson, Merck KGaA, Novartis, Novo Nordisk, Pfizer, Roche, Sanofi, Shionogi, and UCB Pharma (“Data Providers”). Neither TransCelerate Biopharma Inc. nor the Data Providers have contributed to or approved or are in any way responsible for this research result.

References

Alt EM et al. 2024. Leap: the latent exchangeability prior for borrowing information from historical data. Biometrics. 80:ujae083.

Baron E, Zhu J, Tang R, Chen M-H. 2022. Bayesian divide-and-conquer propensity score based approaches for leveraging real world data in single arm clinical trials. J Biopharm Stat. 32:75–89.

Bi D, Ji Y. 2023. A class of dependent random distributions based on atom skipping. arXiv preprint arXiv:2304.14954. preprint: not peer reviewed.

Bi D, Liu M, Lin J, Liu R. 2023. Beats: Bayesian hybrid design with flexible sample size adaptation for time-to-event endpoints. Stat Med. 42:5708–5722.

Chandra NK, Sarkar A, de Groot JF, Yuan Y, Müller P. 2023. Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials. J Am Stat Assoc. 118:2301–2314.

Chen W-C et al. 2020. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. J Biopharm Stat. 30:508–520.

Chipman HA, George EI, McCulloch RE. 2010. Bart: Bayesian additive regression trees.

FDA, US. (2023). Considerations for the design and conduct of externally controlled trials for drug and biological products.

Gelfand AE, Kottas A. 2002. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. J Comput Graph Stat. 11:289–305.

Hey SP, Kimmelman J. 2014. The questionable use of unequal allocation in confirmatory trials. Neurology. 82:77–79.

Hobbs BP, Sargent DJ, Carlin BP. 2012. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. 7:639–674.

Horn RA, Johnson CR. 1990. Norms for vectors and matrices. Matrix analysis, 313–386.

Hubert L, Arabie P. 1985. Comparing partitions. J Classif. 2:193–218.

Hupf B, Bunn V, Lin J, Dong C. 2021. Bayesian semiparametric meta-analytic-predictive prior for historical control borrowing in clinical trials. Stat Med. 40:3385–3399.

Ibrahim JG, Chen M-H. 2000. Power prior distributions for regression models. Stat Sci. 46–60.

Inman HF, Bradley EL. Jr, 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun Stat Theory Methods. 18:3851–3874.

King G, Nielsen R. 2019. Why propensity scores should not be used for matching. Polit Anal. 27:435–454.

Liu M, Bunn V, Hupf B, Lin J, Lin J. 2021. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. Stat Med. 40:4794–4808.

Lu N et al. 2022. Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources. J Biopharm Stat. 32:158–169.

Meilă M. 2007. Comparing clusterings—an information based distance. J Multivar Anal. 98:873–895.

Pitman J. 2002. Poisson–Dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. Combinator Probab Comp. 11:501–514.

Schmidli H et al. 2014. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 70:1023–1032.

Schwartz DE, Ji Y and Wang L 2024. Dynamic borrowing from historical controls via the synthetic prior with covariates in randomized clinical trials.

Simpson EL, others et al. 2022. Efficacy and safety of upadacitinib in patients with moderate to severe atopic dermatitis: analysis of follow-up data from the measure up 1 and measure up 2 randomized clinical trials. JAMA Dermatol. 158:404–413.

Teh Y, Jordan M, Beal M, Blei D. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. Advances in Neural Information Processing Systems 17.

US FDA. (2018). Framework for FDA's real-world evidence program.

Wang C, Chen W-C. 2022. psrwe: Ps-integrated methods for incorporating rwe in clinical studies. r package, version 3.1.

Zhao Z. 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and monte carlo evidence. Rev Econ Stat. 86:91–107.

Zhou T, Ji Y. 2021. Incorporating external data into the analysis of clinical trials via bayesian additive regression trees. Stat Med. 40:6421–6442.