Check for updates

Received: 12 July 2023 | Revised: 7 April 2024 | Accepted: 10 June 2024

DOI: 10.1002/sim.10158

RESEARCH ARTICLE

Statistics in Medicine WILEY

Propensity score weighted multi-source exchangeability models for incorporating external control data in randomized clinical trials

Wei Wei[1] | **Yunxuan Zhang**[1] | **Satrajit Roychoudhury**[2] | **the Alzheimer's Disease Neuroimaging Initiative**[1]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut,

[2]Global Biometrics & Data Management, Pfizer Inc, New York, New York, USA

Correspondence
Wei Wei, Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA.
Email: wei.wei@yale.edu

Funding information
National Cancer Institute, Grant/Award Number: P30-CA016359; National Institute of Dental and Craniofacial Research, Grant/Award Number: P50DE030707

Among clinical trialists, there has been a growing interest in using external data to improve decision-making and accelerate drug development in randomized clinical trials (RCTs). Here we propose a novel approach that combines the propensity score weighting (PW) and the multi-source exchangeability modelling (MEM) approaches to augment the control arm of a RCT in the rare disease setting. First, propensity score weighting is used to construct weighted external controls that have similar observed pre-treatment characteristics as the current trial population. Next, the MEM approach evaluates the similarity in outcome distributions between the weighted external controls and the concurrent control arm. The amount of external data we borrow is determined by the similarities in pretreatment characteristics and outcome distributions. The proposed approach can be applied to binary, continuous and count data. We evaluate the performance of the proposed PW-MEM method and several competing approaches based on simulation and re-sampling studies. Our results show that the PW-MEM approach improves the precision of treatment effect estimates while reducing the biases associated with borrowing data from external sources.

KEYWORDS
Bayesian design, hybrid control, pediatric study, rare disease

1 | INTRODUCTION

Randomized controlled trials (RCTs) are commonly considered as the gold standard for demonstrating the causal effect of an experimental treatment on clinical outcomes of interest. Despite its scientific rigor, the implementation of RCTs can be challenging due to the associated large sample size, long duration, and operational cost. This is specially challenging in rare disease trials and in situations when patients are unwilling to be randomized to a standard of care (SOC) that causes low clinical benefit and/or severe toxicity. To overcome these hurdles, the FDA and the pharmaceutical industry have expressed a growing interest in harnessing external data sources in drug development.[1-6] Hybrid controlled trial (HCT) designs can be constructed by augmenting the internal control arm (IC) of an RCT using patient-level external control (EC) data from prior clinical trials and real world data. The use of HCTs can alleviate the challenges associated with RCTs by reducing the sample size of the IC arm and potentially provide increased statistical power and precision in treatment evaluation.

 © 2024 John Wiley & Sons Ltd.

The selection of external control data requires care and should follow recommendations for systematic reviews.[7,8] This minimizes the risk of systematic biases, which can arise as a result of, for example, changes in standard of care over time, differences in inclusion/exclusion criteria, confounding environmental factors, or the evolution of diagnostic tools. The relevance of external control data depends on a number of factors such as the heterogeneity and natural history of the disease, and requires case-by-case assessment. When suitable external controls are identified, careful statistical considerations are needed to increase the likelihood of distinguishing the effect of a drug from other factors that can confound the outcome of interest.

A number of Bayesian approaches have been developed to incorporate relevant information from external data sources in the form of informative priors, such as the power prior,[9] the modified power prior,[10] the commensurate prior[11] and the meta-analytic-predictive prior (MAP).[12,13] The exchangeability–nonexchangeability (EXNEX) approach can also be used to borrow information from external data sources by constructing a mixture model assuming the parameter of interest in the internal control is either exchangeable or not exchangeable with external data.[14] These approaches provide powerful tools for incorporating external information in design and analysis of clinical trial under Bayesian framework, but they are generally not designed to adjust for pre-treatment covariates and provide causal inference. The majority of statistical approaches for designing HCTs rely on the use of propensity score (PS) based methods to reduce biases resulting from unbalanced covariates distributions between internal and external patient populations. Ventz et al considered the use of external data in making early stopping decisions in RCTs and developed a measure of disimilarity for deciding if external data should be used in evaluating efficacy.[15,16] On the other hand, the propensity score-integrated composite likelihood (PSCL) method[17] takes the divide-and-conquer approach to reduce the heterogeneity between different sources. It first groups patients into PS defined strata and then conducts information borrowing by down-weighting EC patients in each stratum. Within each stratum, IC and EC patients are expected to have similar pre-treatment characteristics and borrowing from the external data can be better justified. The down-weighting step helps to make sure the study result is not dominated by external data and more data is borrowed from homogeneous strata as opposed to heterogeneous strata with different PS distributions. Then, the stratum specific estimates are combined to obtain an overall estimate of treatment effect. Similar strategies have been proposed by applying the MAP method to each stratum to incorporate information from external sources.[18,19] Alternatively, one can use propensity score matching to choose the external control patients for inclusion into analysis as done by the Roche DLBLC study.[20]

A fundamental assumption in causal inference is ignorability. For a HCT, ignorability implies data source (EC vs. IC) is independent of potential outcomes given the propensity score of being in the current study.[21] The ignorability assumption is valid if all confounders have been measured and specified correctly in the PS model. In practice, important prognostic variables such as certain biomarkers are often not present in external data and the PS model might be mis-specified. As a result, statistical inference based on the propensity score processed data might produce misleading estimates if important prognostic variables are left out or not correctly adjusted for in the propensity score model. The PSCL approach evaluates the relevance of external data according to the overlap of propensity score distributions between different data sources, but it doesn't take into account the possibility of unmeasured confounders. Moreover, Bayesian approaches relying on the construction of PS defined strata requires a sufficiently large sample size for both internal and external data sources, which may not be feasible in rare disease trials. Therefore, novel statistical approaches that can better suit the sample size constraint of rare disease trials and provide some safe guard to the biases caused by unknown confounders are urgently needed.

The rest of the paper is organized as follows. We introduce the proposed method for designing HCTs with binary and continuous endpoints in Section 2. In Sections 3 and 4, we evaluate the frequentist operating characteristics of the proposed approach using simulations and a collection of datasets from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). We compare the proposed method with several competing approaches on the basis of type I error rate, power, bias and coverage probability. We discuss its benefits and limitations in Section 5.

2 | METHOD

External control (EC) data can be gathered from different sources including registries, clinical databases or completed and ongoing trials. We assume all EC patients are treated by the same standard of care (SOC) as patients randomized to the IC arm and the same eligibility criteria had been applied to IC and EC patients.

We propose a two step procedure for designing HCTs based on propensity score weighting (PW) and the multi-source exchangeability modelling approach (MEM). Hereafter, we will refer to the proposed approach as PW-MEM. The first step

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

involves standardizing the IC and ECs to take into account the possible imbalance of important prognostic factors. This can be achieved by re-weighting the EC patients using a propensity score model.[22] Assuming the PS model of being in the current study is correctly specified and includes all important baseline covariates, patients from the IC and the propensity score weighted external control (PW-EC) can be considered as random samples from the same population. We assume their clinical outcomes are exchangeable. Following previous works on MEM,[23-26] here we consider the outcome data from IC and PW-EC exchangeable if they are identically distributed, which differs from De Finetti's classical definition of exchangeable random variables. The second step uses a MEM approach proposed by Kaizer et al that provides a statistical framework for evaluating the exchangeability between outcome distributions from different data sources.[23] We will assess the exchangeability in outcome distribution between IC and PW-EC using MEM and borrow data from PW-EC according to its exchangeability with IC.

Compared to the divide-and-conquer approaches such as PSCL and PS-MAP,[17,18] the PW-MEM method doesn't require additional statistical considerations involved in combining stratum-specific estimates. Unlike the divide-and-conquer approaches, which require an adequate sample size for both the IC and EC to fill all the strata, the use of PW allows us to work with studies of relatively small sample sizes typically seen in rare disease research. Compared to the previous MEM approach based on Bayesian model averaging,[23] the proposed PW-MEM approach allows us to calculate the sample size incorporated from external sources conveniently by counting each external patient as a fraction of an internal patient. Furthermore, the PW-MEM approach can be generalized to situations when multiple external control datasets are available.

2.1 | Notations

Consider a clinical trial randomizing patients to either the internal control arm (IC) or the internal treatment arm (IT). The resulting sample size for IC and IT arm is n_0 and n_T, respectively. Let θ_0 and θ_T denote the parameter of interest (response rate, mean blood pressure, etc.) in the IC and IT arm. Denote Δ the treatment effect expressed as a contrast between these two arms (eg, $\Delta = \theta_T - \theta_0$). In this work, we are interested in improving the precision in estimating Δ by leveraging data from external controls.

Let D_j denote a data source, with $D_j = 0$ representing the current control and $D_j = 1, \ldots, J$ representing the J chosen external controls. Denote n_j the sample size available from D_j. Let i be the index of a patient from one of the $J + 1$ sources. Denote y_i and S_i the clinical outcome and the indicator of IC versus EC for patient i. We have $S_i = 1$ if this patient is from D_0 and $S_i = 0$ if this patient is from external sources. Let $\mathbf{X}_i$ represent a length q vector of pre-treatment variables for the same patient. We assume these pre-treatment variables are commonly available in different data sources. We will construct the hybrid control (HC) using data from $D_0, \ldots, D_J$. The HC consists of $n_{HC} = \sum_{j=0}^{J} n_j$ patients.

2.2 | Propensity score weighting

Denote e_i as the propensity score of patient i conditional on $\mathbf{X}_i$, for $i = 1, \ldots, n_{HC}$. We define e_i as the conditional probability of a patient being in the internal study ($e_i = Prob(S_i = 1|\mathbf{X_i})$), which is commonly estimated using a logistic regression. Though we focus on using logistic regression in this work for illustration purposes, it should be noted that several machine-learning methods can also be used.[27-29] As our objective is to incorporate external data into the current study, we will use IC patients as the target population to which the EC patients is standardized. We define the propensity score weight as

$$w_i = S_i + \frac{e_i(1 - S_i)}{1 - e_i}. \tag{1}$$

Based on this choice of weight, external patients ($S_i = 0$) are weighted as $w_i = \frac{e_i}{1-e_i}$, whereas internal patients ($S_i = 1$) have a constant weight of $w_i = 1$. Borrowing data from PW-EC is less likely to introduce bias because of the similarity in baseline covariates between the PW-EC and IC patients. However, bias can still be introduced due to unmeasured confounders left out by the propensity score model. The following section introduces the proposed PW-MEM method for leveraging external data based on propensity score weighting and the multisource exchangeability modelling (MEM) approach.

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.3 | Propensity score weighted multi-source exchangeability modelling

Let M_k denote a specific exchangeability configuration for the weighted outcome data from IC and ECs. Under M_k, these $J+1$ control groups can be classified into P_k clusters denoted as $M_{k,1}, \ldots, M_{k,P_k}$. Let $\theta_1, \ldots, \theta_J$ denote the parameter of interest (response rate, mean glucose level, etc) for the J external controls after propensity score weighting. We assume the weighted outcome data from sources located in the same cluster $M_{k,p}$ are exchangeable and have the same parameter $\theta_{k,p}$ (ie, $\theta_j = \theta_{k,p}$ if $j \in M_{k,p}$), where $p = 1, \ldots, P_k$ and $j = 0, \ldots, J$. The number of distinct parameters in M_k is P_k.

Consider the case when there exists two ECs, there are five possible exchangeability patterns:

$$
\begin{aligned}
M_1&: \theta_0 = \theta_1 = \theta_2,\\
M_2&: \theta_0 = \theta_2 \neq \theta_1,\\
M_3&: \theta_0 = \theta_1 \neq \theta_2,\\
M_4&: \theta_1 = \theta_2 \neq \theta_0,\\
M_5&: \theta_0 \neq \theta_1 \neq \theta_2.
\end{aligned}
$$

where $P_1 = 1$, $P_2 = P_3 = P_4 = 2$, and $P_5 = 3$.

If M_1 is true, then the outcome distributions in the IC and the two PW-ECs are identical, suggesting PS weighting is likely to have eliminated all the possible confounders. In this case, we can conduct a pooled analysis by combining weighted outcome data from different sources. If M_4 or M_5 is true, the outcome distributions are different between the IC and the PW-ECs, suggesting the presence of residual confounding and we should not borrow information from the external controls. Under other exchangeability patterns, we may only want to borrow data from certain external sources. To make a better informed decision about the amount of borrowing, we will need to estimate the uncertainty of these different exchangeability configurations.

Given the estimated weights $\mathbf{w} = (w_1, \ldots, w_{n_{HC}})$, the weighted likelihood function for outcome data in cluster $M_{k,p}$ is defined as

$$
\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}|M_{k,p}, \theta_{k,p}) = \prod_{j \in M_{k,p}} L_{\mathbf{w}}(\mathbf{y}_j|\theta_{k,p}), \tag{2}
$$

where $\mathbf{y}_j$ is the collection of outcome data from D_j. We define

$$
L_{\mathbf{w}}(\mathbf{y}_j|\theta_{k,p}) = \prod_{i \in D_j} f(y_i|\theta_{k,p})^{w_i},
$$

with $f(.)$ denoting the probability density or probability mass function.

Assume the prior density for $\theta_{k,p}$ is $\pi(\theta_{k,p})$. The weighted marginal likelihood for outcome data in $M_{k,p}$ can be calculated by averaging the weighted likelihood for data in this cluster over the prior density of $\theta_{k,p}$:

$$
\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}|M_{k,p}) = \int \mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}|M_{k,p}, \theta_{k,p})\pi(\theta_{k,p})d\theta_{k,p}. \tag{3}
$$

Generally, this integral can be evaluated numerically using the *Integrate* function in R base package. For several commonly used distributions (binomial, Gaussian, Poisson), the weighted marginal likelihood can be found in close form. We summarize the weighted marginal likelihood function and associated assumptions for these models in Table 1.

Denote $\pi(M_k)$ the prior probability of M_k. Let K be the number of possible exchangeability configurations. Following Bayes rule, we can estimate the posterior probability of M_k as:

$$
\pi(M_k|\mathbf{w}, \mathbf{y}_0, \ldots, \mathbf{y}_J) = \frac{\pi(M_k)\prod_{p=1}^{P_k}\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}|M_{k,p})}{\sum_{k'=1}^{K}\pi(M_{k'})\prod_{p=1}^{P_{k'}}\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k',p}|M_{k',p})}. \tag{4}
$$

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Marginal likelihood for data in cluster $M_{k,p}$ given the propensity score-based-weights, assuming the outcome distribution is Gaussian, Binomial or Poisson.

Normally distributed data	
Weighted Likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}\|\theta_{k,p}, M_{k,p}) = \prod_{j\in M_{k,p}}\prod_{i\in D_j}\left\{\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left[-\frac{1}{2}\left(\frac{y_i-\theta_{k,p}}{\sigma_j}\right)^2\right]\right\}^{w_i}$
Prior	$\pi(\theta_{k,p}) \propto 1$
Marginal likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}\|M_{k,p}) \propto exp(\frac{u^2}{4v})\sqrt{\frac{\pi}{v}}$
Notes	$u = \sum_{j\in M_{k,p}}\sum_{i\in D_j}\frac{w_i y_i}{\sigma_j^2}, \quad v = \frac{1}{2}\sum_{j\in M_{k,p}}\sum_{i\in D_j}\frac{w_i}{\sigma_j^2}$ $\sigma_j^2 = \frac{\sum_{i\in D_j} w_i(y_i-\bar{y}_j)^2}{\sum_{i\in D_j} w_i}, \quad \bar{y}_j = \frac{\sum_{i\in D_j} w_i y_i}{\sum_{i\in D_j} w_i}$
Binary data	
Weighted Likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}\|M_{k,p}, \theta_{k,p}) \propto \prod_{j\in M_{k,p}}\prod_{i\in D_j}\left\{\theta_{k,p}^{y_i}(1-\theta_{k,p})^{1-y_i}\right\}^{w_i}$
Prior	$\pi(\theta_{k,p}) = Beta(\alpha, \beta)$
Marginal likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}\|M_{k,p}) \propto \frac{B(\alpha+\sum_{j\in M_{k,p}} r_j^w,\ \beta+\sum_{j\in M_{k,p}}(n_j^w-r_j^w))}{B(\alpha,\beta)}$
Notes	$n_j^w = \sum_{i\in D_j} w_i, \quad r_j^w = \sum_{i\in D_j} w_i y_i$
Poisson data	
Weighted Likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}, \|M_{k,p}, \theta_{k,p}) \propto \prod_{j\in M_{k,p}}\prod_{i\in D_j}\left\{e^{-\theta_{k,p}}\theta_{k,p}^{y_i}\right\}^{w_i}.$
Prior	$\pi(\theta_{k,p}) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta_{k,p}^{\alpha-1}e^{-\beta\theta_{k,p}}$
Marginal Likelihood	$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_{k,p}\|M_{k,p}) \propto \frac{\Gamma(\alpha+\sum_{j\in M_{k,p}}\sum_{i\in D_j} w_i y_i)}{(\beta+\sum_{j\in M_{k,p}}\sum_{i\in D_j} w_i)^{\alpha+\sum_{j\in M_{k,p}}\sum_{i\in D_j} w_i y_i}}$

Note: We assume the variance is known under the Gaussian case.

Let λ_j denote the probability of $\theta_0 = \theta_j$, which represents the uncertainty about the exchangeable assumption between the j-th PW-EC and the IC. We can estimate λ_j by summing up the posterior probabilities of exchangeability configurations under which $\theta_j = \theta_0$, which is expressed as

$$\lambda_j = \sum_{k=1}^{K}\pi(M_k|\mathbf{w}, \mathbf{y}_0, \ldots, \mathbf{y}_J)I(\theta_j = \theta_0|M_k), \tag{5}$$

where $I(.)$ is the indicator function.

2.4 | Posterior inference

We will conduct pooling according to the exchangeability between each weighted external control and the internal control. Specifically, we will pool IC with an PW-EC if they are considered exchangeable and the amount of pooling will be determined by λ_j, which takes into account the uncertainty related to the exchangeable assumption. Assume $\theta_0 = \theta_1 = \cdots \theta_J$, we define the weighted likelihood function for outcome data from $D_0, \ldots, D_J$ as

$$\mathbb{L}_{\mathbf{w}}(\mathbf{y}_0, \ldots, \mathbf{y}_J|\theta_0, \lambda_1, \ldots, \lambda_J) = \mathbb{L}(\mathbf{y}_0|\theta_0)\left\{\prod_{j=1}^{J}\mathbb{L}_{\mathbf{w}}(\mathbf{y}_j|\theta_0)^{\lambda_j}\right\} = \mathbb{L}(\mathbf{y}_0|\theta_0)\left\{\prod_{j=1}^{J}\prod_{i\in D_j} f(y_i|\theta_0)^{w_i\lambda_j}\right\}. \tag{6}$$

Assume the prior distribution for θ_0 is $\pi(\theta_0)$. The posterior distribution of θ_0 is

$$\pi(\theta_0|\mathbf{y}_0, \ldots, \mathbf{y}_J, \lambda_1, \ldots, \lambda_J, \mathbf{w}) \propto \mathbb{L}_{\mathbf{w}}(\mathbf{y}_0, \ldots, \mathbf{y}_J|\theta_0, \lambda_1, \ldots, \lambda_J)\pi(\theta_0). \tag{7}$$

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

For binary outcomes, let $r_0 = \sum_{i \in D_0} y_i$ denote the number of successes out of the n_0 patients in the internal control. Assume the prior distribution for θ_0 is $Beta(\alpha, \beta)$, the posterior density of θ_0 given $\lambda_1, \ldots, \lambda_J$ and $\mathbf{w}$ is

$$\pi(\theta_0 | \mathbf{y}_0, \ldots, \mathbf{y}_J, \lambda_1, \ldots, \lambda_J, \mathbf{w}) = Beta(\alpha + r_0 + \sum_{j=1}^{J} \lambda_j r_j^w,\ \beta + n_0 - r_0 + \sum_{j=1}^{J} \lambda_j (n_j^w - r_j^w)),$$

where $n_j^w = \sum_{i \in D_j} w_i$, and $r_j^w = \sum_{i \in D_j} w_i y_i$.

For normally distributed outcomes, if we assume the prior density for θ_0 is $\pi(\theta_0) \propto 1$, then the posterior density of θ_0 conditional on $\lambda_1, \ldots, \lambda_J$ and $\mathbf{w}$ is normally distributed with mean μ and precision τ, where

$$\tau = \frac{n_0}{\sigma_0^2} + \sum_{j=1}^{J} \frac{\lambda_j n_j^w}{\sigma_j^2}$$

and

$$\mu = \frac{1}{\tau} \left\{ \frac{n_0 \bar{y}_0}{\sigma_0^2} + \sum_{j=1}^{J} \frac{\lambda_j n_j^w \bar{y}_j}{\sigma_j^2} \right\}.$$

We consider $\sigma_0^2 = \frac{\sum_{i \in D_0} (y_i - \bar{y}_0)^2}{n_0}$ and $\sigma_j^2 = \frac{\sum_{i \in D_j} w_i (y_i - \bar{y}_j)^2}{\sum_{i \in D_j} w_i}$ for $j = 1, \ldots, J$. Likewise, $\bar{y}_j$ is the weighted average of outcome data in D_j.

2.5 | Effective sample size of external data

A key question in assessing the impact of external data on the hybrid control is to determine the amount of data that has actually been incorporated from external sources. The PW-MEM approach allows us to compute the effective sample size (ESS) we incorporate from external controls conveniently.

The PW-MEM method relies on the availability of both internal and external data for the estimation of propensity score and exchangeability between different data sources. Based on equation 6, an external patient from D_j is considered as equivalent to $\lambda_j w_i$ patient in the IC after PS weighting, where w_i is the PS-based weight for this patient and λ_j is the posterior probability that the parameter of interest from an external source is exchangeable with the internal source (ie, $\theta_0 = \theta_j$). Thus, the ESS we incorporate from all the J external sources is

$$ESS = \sum_{j=1}^{J} \sum_{i \in D_j} \lambda_j w_i.$$

PS weighting produces a more homogeneous population with balanced pretreatment covariates between internal and external data sources. Multisource exchangeability modelling helps to safeguard the hybrid controlled design from biases introduced by residual confounding after PS weighting. We will leverage more data from an external source if its outcome distribution after PS weighting is similar to IC. Likewise, we will have less data from an external source if its outcome distribution is still considerably different from IC after PS weighting.

3 | SIMULATION STUDIES FOR BINARY OUTCOMES

In this section, we conduct simulation studies based on the MORPHEUS-CRC trial, which is a phase Ib/II randomized controlled trial investigating the effect of a novel immune combination compared to regorafenib in metastatic colorectal cancer patients.[20] Assume 90 patients will be randomized in 2:1 ratio to either the experimental ($n_T = 60$) or the control arm ($n_0 = 30$). Suppose there exists an EC cohort consisting of 1000 patients. We assume five binary pre-treatment variables ($X_1, \ldots, X_5$) and one normally distributed pre-treatment variable (X_6) are commonly available in the internal and

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 The distribution of pre-treatment covariates under different simulation scenarios for a hypothetical trial with binary outcomes.

	Proportion of pre-treatment covariates in IC(EC)					Intercept		Response rate	
Scenario	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	$X_4 = 1$	$X_5 = 1$	**IC**	**EC**	**IC**	**EC**
1	0.43 (0.43)	0.84 (0.84)	0.54 (0.54)	0.62 (0.62)	0.64 (0.64)	−2	−2	0.20	0.20
2	0.43 (0.43)	0.84 (0.84)	0.54 (0.54)	0.62 (0.62)	**0.20 (0.64)**	−2	−2	0.27	0.20
3	0.43 (0.43)	0.84 (0.84)	0.54 (0.54)	0.62 (0.62)	**0.90 (0.64)**	−2	−2	0.16	0.20
4	0.43 (0.43)	0.84 (0.84)	0.54 (0.54)	0.62 (0.62)	0.64 (0.64)	**−1.5**	−2	0.29	0.20

Note: There is no confounder in Scenario 1. There exists one confounder (X_5) in Scenarios 2 and 3. There exists an unobserved confounder in Scenario 4, which is represented a different intercept term. The marginal response rates for EC and IC are shown for each scenario. The coefficients of $X_1, \ldots, X_5$ are fixed across scenarios.

external datasets. We generate $\mathbf{X} = (X_1, \ldots, X_6)$ independently of each other for this trial and the external dataset, respectively. Let $\mathbf{b} = (b_1, \ldots, b_6)$ represent the effect of pre-treatment variables on the outcome y, which indicates whether a patient responds to treatment. In this case, θ_0 and θ_T represent the probability of response for the internal control and experimental arm, respectively. Here Δ denote the log odds ratio of the treatment effect. We assume $\Delta = 0$ under the null and $\Delta = 1$ under the alternative. Denote b_{0S} the intercept for data source S. We simulate the binary outcomes y assuming

$$P(y = 1|\mathbf{X}, T, S) = \frac{exp(b_{0S} + X_1 b_1 + \cdots + X_6 b_6 + T\Delta)}{1 + exp(b_{0S} + X_1 b_1 + \cdots + X_6 b_6 + T\Delta)},$$

where T is the indicator of treatment assignment.

We consider four simulation scenarios and generate 5,000 trials in each scenario under the null and the alternative, respectively. Table 2 summarizes the distribution of pre-treatment variables and the marginal probability of response associated with EC and IC in each scenario (i.e., θ_0 and θ_T). We assume $\mathbf{b} = (0.2, 0.2, 0.5, 1, -1, 0)$ is fixed across scenarios. The intercept term b_{0S} is allowed to vary across scenarios. We assume the intercept term for IC(EC) is $-2(-2)$, $-2(-2)$, $-2(-2)$, and $-1.5(-2)$ in scenario $1, \ldots, 4$, respectively.

In Scenario 1, the distribution of pre-treatment variables are identical between EC and IC. Senario 1 represent situations when the pre-treatment covariates are balanced between the external data and internal data. In Scenarios 2 and 3, the distribution of prognostic variable X_5 is different between IC and EC. Thus, X_5 is the confounder in these scenarios. In Scenario 4, the distribution of $X_1, \ldots, X_5$ are identical between IC and EC, but there exists an unobserved confounder, which is represented by the different intercept terms for IC and EC.

We evaluate treatment effect in each simulated trial using the following approaches:

1. IN-test: only use IC data assuming a $Beta(0.5, 0.5)$ prior for θ_0.
2. Pooled-test: combine data from IC and EC assuming $\pi(\theta_0) \sim Beta(0.5, 0.5)$.
3. Construct a hybrid control using PW-MEM.
4. Construct a hybrid control using rMAP.
5. Apply the marginal structure model (MSM) with a logit link to the combined data, adjusting the effect of pre-treatment variables.

When there is only one EC, there exists two exchangeability configurations: $M_1 : \theta_0 = \theta_1$ vs. $M_2 : \theta_0 \neq \theta_1$. We should pool EC and IC together if M_1 is true and we should only rely on IC if M_2 is true. For PW-MEM, we assume the prior probability of pooling and not pooling is equally likely, that is, $\pi(M_1) = \pi(M_2) = 0.5$. We will evaluate the effect of different prior choices for PW-MEM in Section 3.2.

For the rMAP approach, we construct an informative prior for θ_0 based on EC data using the meta-analytic-predictive prior (MAP) method. To protect against type-I error rate inflation in the presence of prior-data conflict, the MAP is robustified by adding a non-informative component corresponding to data from one observation. The weight of the robust component is set to be 0.5. We set up a grid of values representing different levels of between-trial heterogeneity and calculate the prior ESS of the rMAP prior based on each value in the grid. The rMAP prior is calibrated by selecting the value of between-trial heterogeneity such that the prior ESS approximately equals to the sample size of IC to prevent excessive

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

borrowing. Standard Bayesian calculus for mixture models implies that the posterior distribution is again a mixture, with component-wise posterior distributions and updated mixture weights.[30,31] The latter depend on the a-priori weights and on how likely the data are under the mixture components. We implement the rMAP approach using the RBest package.[32]

For method (1)–(4), we will conclude the treatment is effective if the posterior probability of $\theta_T > \theta_0$ is greater than a prespecified threshold ϕ, assuming a $Beta(0.5, 0.5)$ prior for θ_T and $\phi = 0.95$.

For MSM, we assume a one-sided significance level of 0.05 and conduct hypothesis testing based on robust standard error. MSM and PW-MEM are based on the same weight function defined by Equation (1).

The R code for implementing the PW-MEM approach can be accessed at https://github.com/smartbenben/PW-MEM.

3.1 | Simulation results

We summarize the performance of different approaches in Figure 1. Under Scenario 1, the distribution of pre-treatment variables are identical between different data sources. In this case, statistical approaches capable of incorporating EC data (MSM, PW-MEM, rMAP and pooled-test) drastically boost the statistical power in this scenario without inflating the type I error rate.

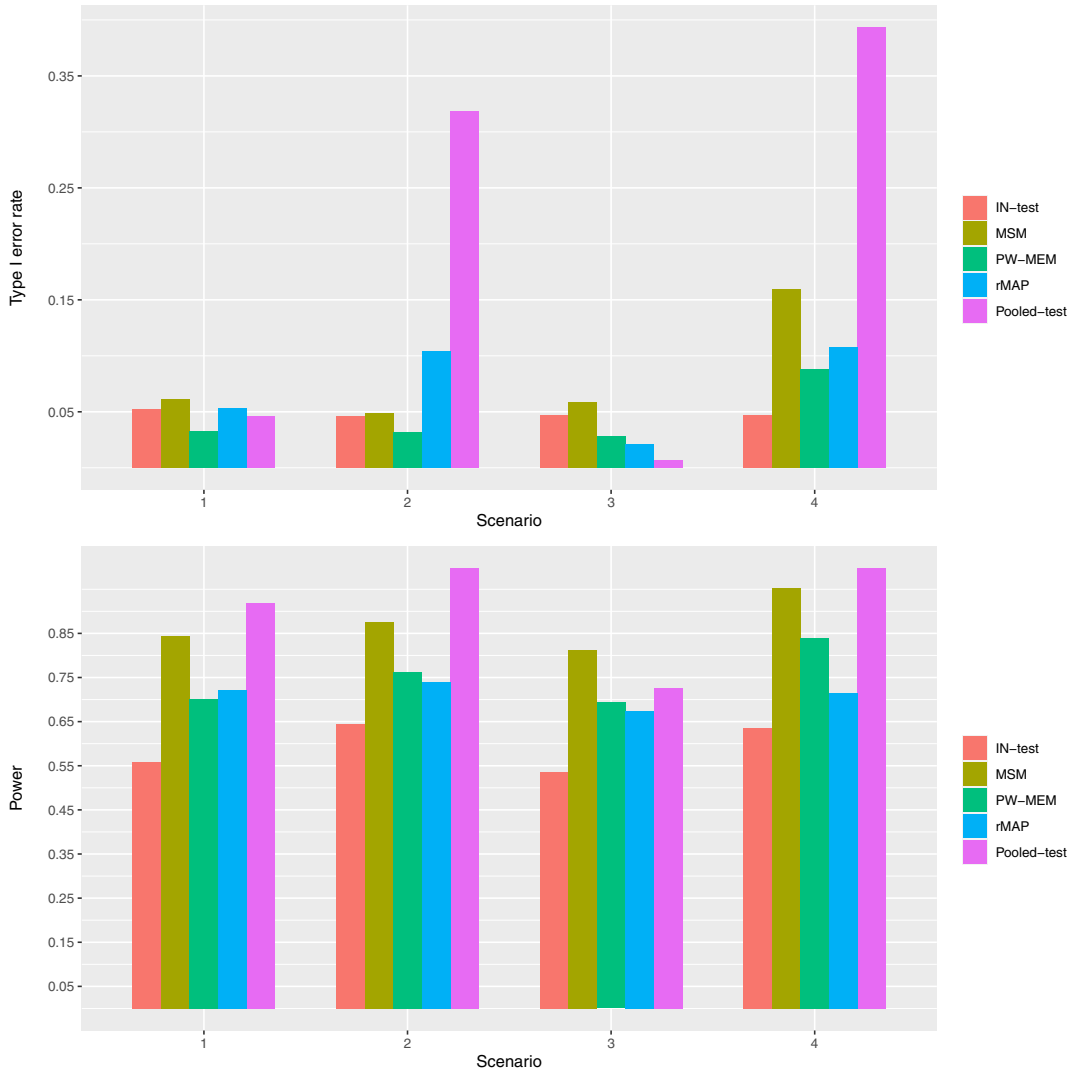

FIGURE 1 The probability of rejecting the null using different approaches under the null and the alternative based on simulation studies assuming the clinical outcomes are binary.

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In Scenarios 2 and 3, the distribution of prognostic variable X_5 differs between IC and EC, resulting in substantial heterogeneity between these patient populations. In Scenario 2, the marginal probability of response in IC and EC is 0.27 and 0.20, respectively. Simply combining EC and IC data results in a downward bias for the estimated response rate in the control group and the pooled-test produces a type I error rate greater than 0.30. In Scenario 3, the marginal probability of response for IC (0.16) is lower than that of EC (0.20). The pooled-test has lower power than MSM because its has an upward bias in estimating the response rate of control arm. In Scenarios 2 and 3, when there exists no unobserved confounders, marginal structure model based on PW effectively eliminates the bias resulting from the confounding of X_5 and provides more power compared to IN-test with slightly inflated type I error rate. In both scenarios, the proposed PW-MEM method is able to provide more power than the IN-test. It is more conservative compared to MSM as it further down-weights the external data according to external-internal outcome similarities.

The internal-external data conflict present in Scenario 4 is caused by an unknown confounder. The marginal probability of response for IC (0.29) in Scenario 4 is higher than that of EC (0.20). MSM is not effective at reducing the systematic differences between EC and IC in this case because the unknown confounders are not included in the estimation of propensity scores and are not included as covariates. As a result, MSM leads to inflated type I error rate in Scenario 4. Compared to MSM, PW-MEM is able to provide better protection on type I error rate by reducing the influence of EC data according to the similarity of outcome data between EC and IC.

Compared to IN-test, the proposed PW-MEM approach provides more statistical power in all scenarios considered and is able to maintain the type I error rate at target level when all confounders have been accounted for in the PS model. Compared to the MSM approach, PW-MEM can provide better control of type I error rate in the presence of unknown confounders. In comparison with rMAP, PW-MEM provides better control of type I error rate especially under Scenario 2 and achieves higher power in Scenarios 2–4.

3.2 | Prior sensitivity analysis

We assess the sensitivity of the PW-MEM approach to different prior choices by considering $\pi(M_1) = (0.05, 0.33, 0.5, 0.67, 0.95)$ in the aforementioned simulation scenarios. We summarize the results of prior sensitivity analysis in Figure 2. For all the prior choices we considered, their type I error rates are below 0.05 in Scenarios 1–3 and close to 0.10 under Scenario 4 in the presence of unobserved confounders.

In Scenarios 1–3, the PS weighting method is able to eliminate the confounding due to unbalanced pre-treatment covariates between EC and IC. The prior strongly favoring the non-exchangeability of the PW-EC and IC ($\pi(M_1) = 0.05$) has larger type I error rate and smaller power compared to other prior choices. In contrast, prior choices considering the exchangeable assumption more plausible can borrow more data from the PW-EC group, resulting in reduced type I and type II error rates.

In Scenario 4, the PS weighting method is not able to eliminate the bias resulting from unobserved confounders. Regardless the choice of priors, the PW-MEM method leads to inflated type I error rate and priors more reluctant to borrow are associated with smaller type I error rates. Nevertheless, even with the most aggressive prior ($\pi(M_1) = 0.95$), the type I error rate is less than 0.10.

4 | REAL DATA APPLICATION

4.1 | ADNI data

In this section, we will apply the PW-MEM method to a hypothetical hybrid controlled trial with continuous endpoint using data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership with the aim of testing whether serial magnetic resonance imaging (MRI), clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The purpose of this example is to illustrate the utility of the proposed methodology while using HCT for development of new drug in this therapeutic area.

The dataset we obtained from ADNI consists of records from 1467 individuals including patient-level data on demographics (age, gender, years of education), baseline lab tests (ApoE4 status), scores of cognitive tests at baseline and over a period of 18 months. For demonstration purpose, the endpoint we focus on here is the 18-month changes in

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

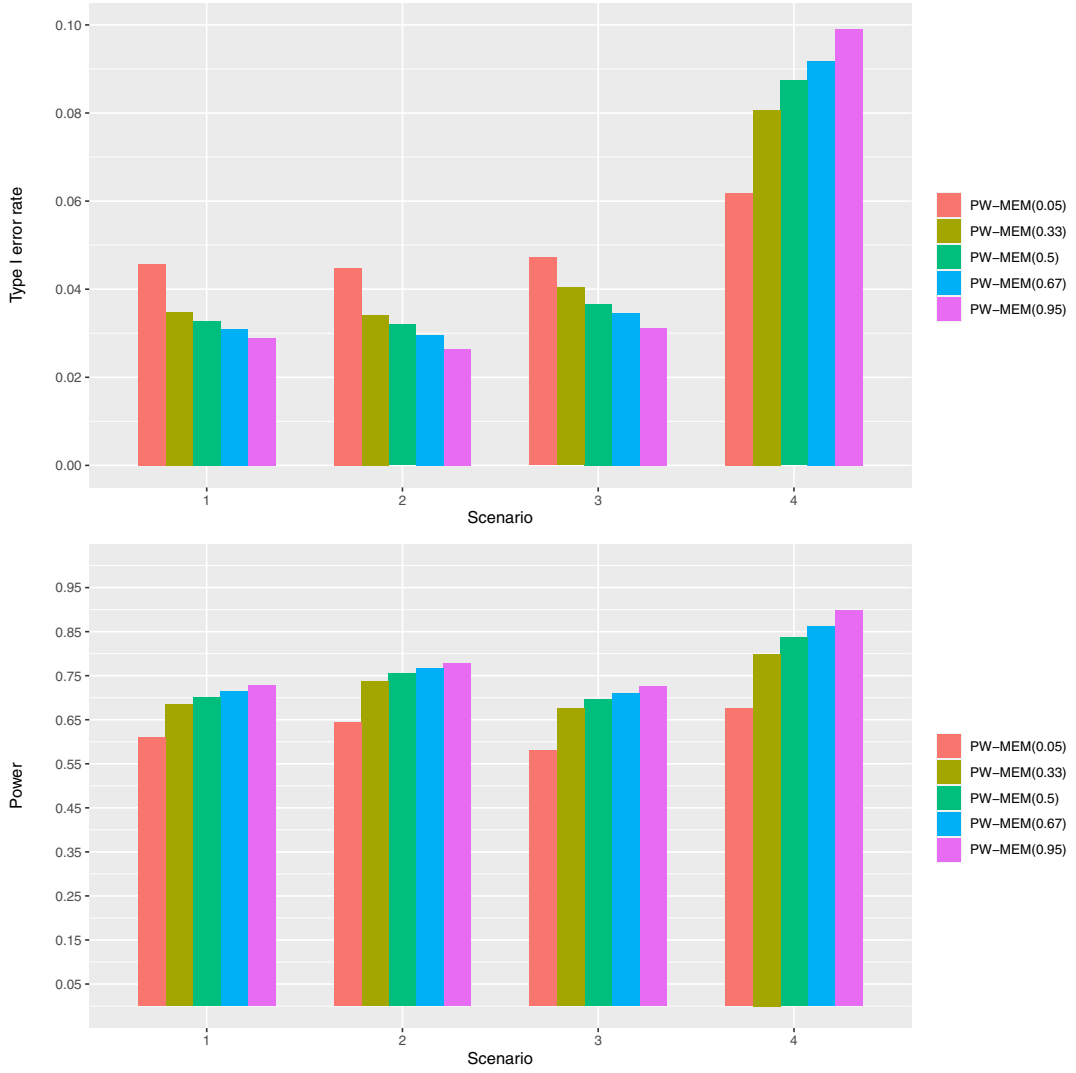

FIGURE 2 The probability of rejecting the null based on different prior choices for PW-MEM under different simulation scenarios, assuming the prior probability that the weighted external control is exchangeable with the internal control is 0.05, 0.33, 0.5, 0.67, and 0.95.

Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog11).[33] We consider external control (EC) as the cohort of patients ($n = 727$) with the date of examination before 2011. This group seems most relevant and comparable with the control arm of current study. Then we evaluate the performance of PW-MEM and several competing approaches as follows:

1. Take a bootstrap sample of 90 patients who are not included in the EC.
2. Randomize these 90 patients in 2:1 ratio to the treatment (IT) and the control group (IC). This is our internal data for the current trial.
3. For patients randomized to the treatment arm, add a treatment effect $\Delta = 0$ under the null hypothesis, and $\Delta = -2$ under the alternative hypothesis.
4. We test treatment effect using four different approaches: (1) test treatment effect based on only the internal data (IN-test); (2) test treatment effect by naively combining internal and external data (pooled-test); (3) Marginal structure model (MSM) adjusting the effect of pre-treatment variables; (4) PW-MEM; (5) An approach based on Bayesian robust meta-analytic-predictive prior (rMAP).
5. Repeat steps 1–4 for 1,000 times.

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Mean differences in baseline covariates between the external and internal control before and after propensity score weighting.

Covariates	Before weighting			After weighting		
	25%	Median	75%	25%	Median	75%
AGE (Years)	1.82	2.72	3.63	−0.03	0.01	0.05
Male (%)	−2.0	5.0	8.0	−0.1	0.1	0.3
Years of education	−0.91	−0.61	−0.24	−0.01	0.00	0.01
APOE4 (%)	−5.0	2.0	9.0	−0.4	−0.1	0.1
ADAS11 at baseline	0.84	1.58	2.35	0.01	0.03	0.07

Note: The mean differences for baseline covariates are estimated in each bootstrap sample and the distribution of mean differences for each covariate are summarized using their median, 25% and 75% percentiles.

We conduct IN-test and pooled-test by regressing the changes in ADAS-Cog11 scores on treatment group. Similar to the previously described simulation studies, we assume the prior probability for the two exchangeable configurations are equally likely ($\pi(M_1) = \pi(M_2) = 0.5$) in setting up PW-MEM. The rMAP prior is constructed to include a robust component corresponding to one observation with a prior weight of 0.5. It should be noted that MSM and PW-MEM are based on the same weight function defined in Equation (1).

We consider a one-sided significance level of 0.05 for IN-test, pooled-test and MSM. For PW-MEM and rMAP, we assume θ_T has a normal prior with a standard deviation of 100, that is, $\pi(\theta_T) \sim N(0,100)$ and conclude the experimental treatment is superior than the control if the posterior probability of having any cognitive improvement given all available data D is greater than a pre-specified threshhold ϕ:

$$P(\Delta < 0|D) > \phi,$$

where $\Delta = \theta_T - \theta_0$ and $\phi = 0.95$.

4.2 | Results

The hypothetical trial based on ADNI data represents a situation when only a limited number of pre-treatment variables are available from the external data and the effectiveness of PS based approaches in eliminating confounders is questionable.

For each bootstrapped sample, we calculate the mean differences of baseline covariates between the external and internal control groups before and after PS weighting. We summarize the distribution of covariate mean differences in Table 3. Based on Table 3, The use of PS weighting effectively eliminates the unbalance in covariates distributions between IC and EC.

We summarize the type I error rate, power and coverage probability of PW-MEM and several competing statistical approaches in Figure 3. A traditional RCT evaluating treatment effect using only internal data (IN-test) has type I error rate and coverage probabilities close to the pre-specified target levels, but it provides only 66.7% power in this hypothetical, 2:1 randomized trial. In contrast, the pooled-test incorporating all EC data without any considerations for external-internal data conflict results in a severe inflation of type I error rate (~ 0.279) and provides low coverage compared to other approaches. Compared to naive pooling approach, the MSM method can offer some protection against internal-external data conflict through PW. However, the usefulness of PW in this case might be limited because only a few pre-treatment covariates are available and it is very likely that important prognostic variables are not included in estimating the propensity scores. As a result, the MSM approach also has inflated type I error rate and reduced coverage probabilities. Compared to MSM, the PW-MEM approach is able to detect the remaining discrepancies between PW-EC and IC data and down-weight the PW-EC data according to their similarity in outcome distributions. Our bootstrap study shows PW-MEM maintains a type I error rate close to the target level of 0.05, achieves more statistical power than IN-test (74.7% vs. 66.7%) and provides coverage probabilities close to 0.95. The rMAP approach performs similarly to PW-MEM, albeit showing inflated type I error rate (~ 0.08).

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

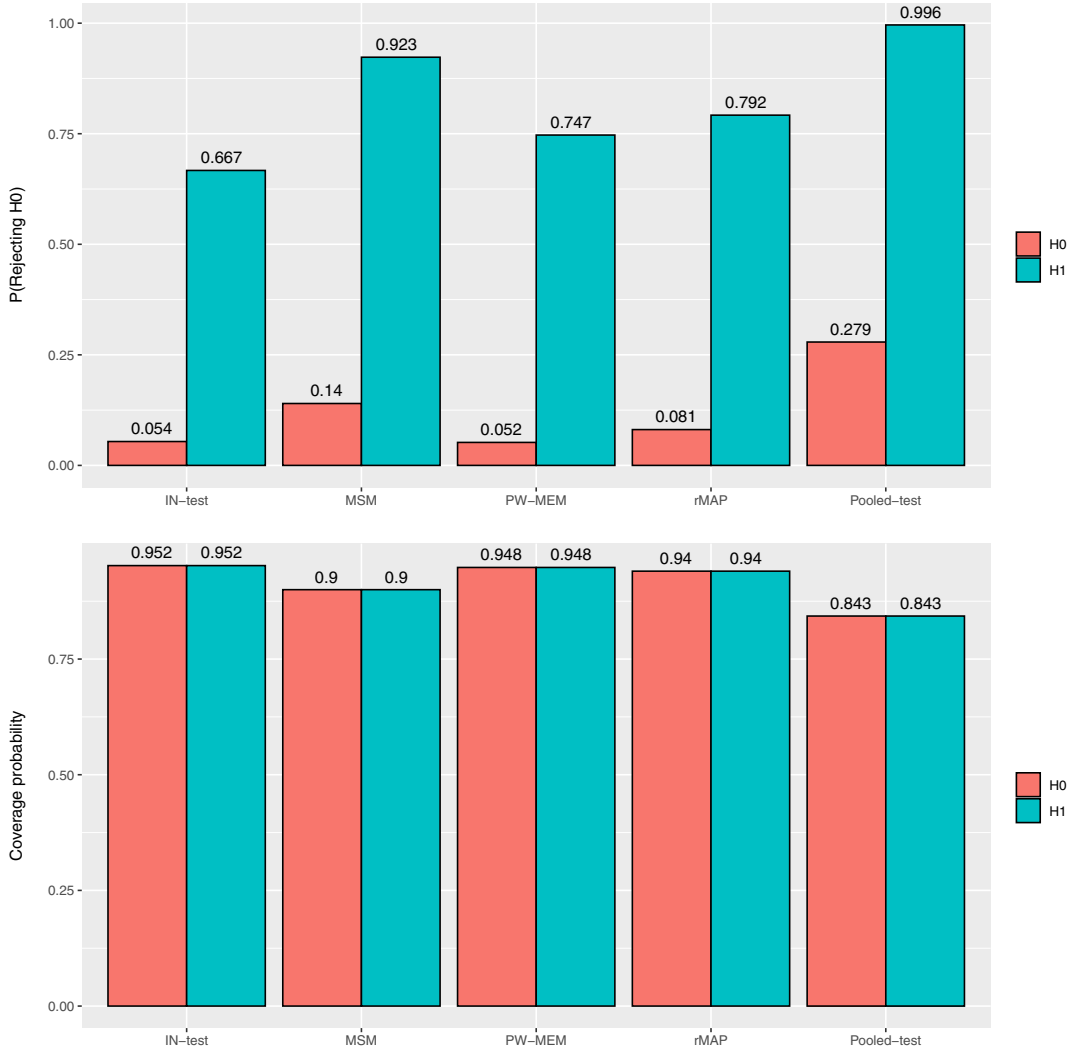

FIGURE 3 The probability of rejecting the null and the coverage probability of different approaches under the null and the alternative hypothesis based on bootstrapping from the ADNI dataset.

Figure 4 summarizes the bias of treatment effect estimates based on these approaches. Compared to MSM and rMAP, PW-MEM is more effective at controlling the bias caused by internal-external heterogeneity.

5 | DISCUSSION

In this work, we develop a novel statistical strategy for augmenting the control arm of RCTs by leveraging external data. The proposed PW-MEM is easy to implement as all the posterior computations can be performed in close form without the need for MCMC. All the parameters in the PW-MEM approach can be specified at the trial designing stage and requires minimal model calibration. Compared to the divide-and-conquer strategies,[17,18,34] PW-MEM can be applied to studies of relatively small sample sizes, which is often seen in trials of rare diseases and oncology. Based on simulation and resampling studies, we demonstrate the use of PW-MEM can provide considerable increases in statistical power and protects the type I error rate from overly inflated in the presence of both known and unknown confounders.

Implementing the PW-MEM approach requires the evaluation of marginal likelihood function. In this work, we focus on single parameter models as their marginal likelihood either exist in close form or can be solved using numerical integration. Expanding the proposed method to more complex models will be more challenging as the marginal likelihood

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

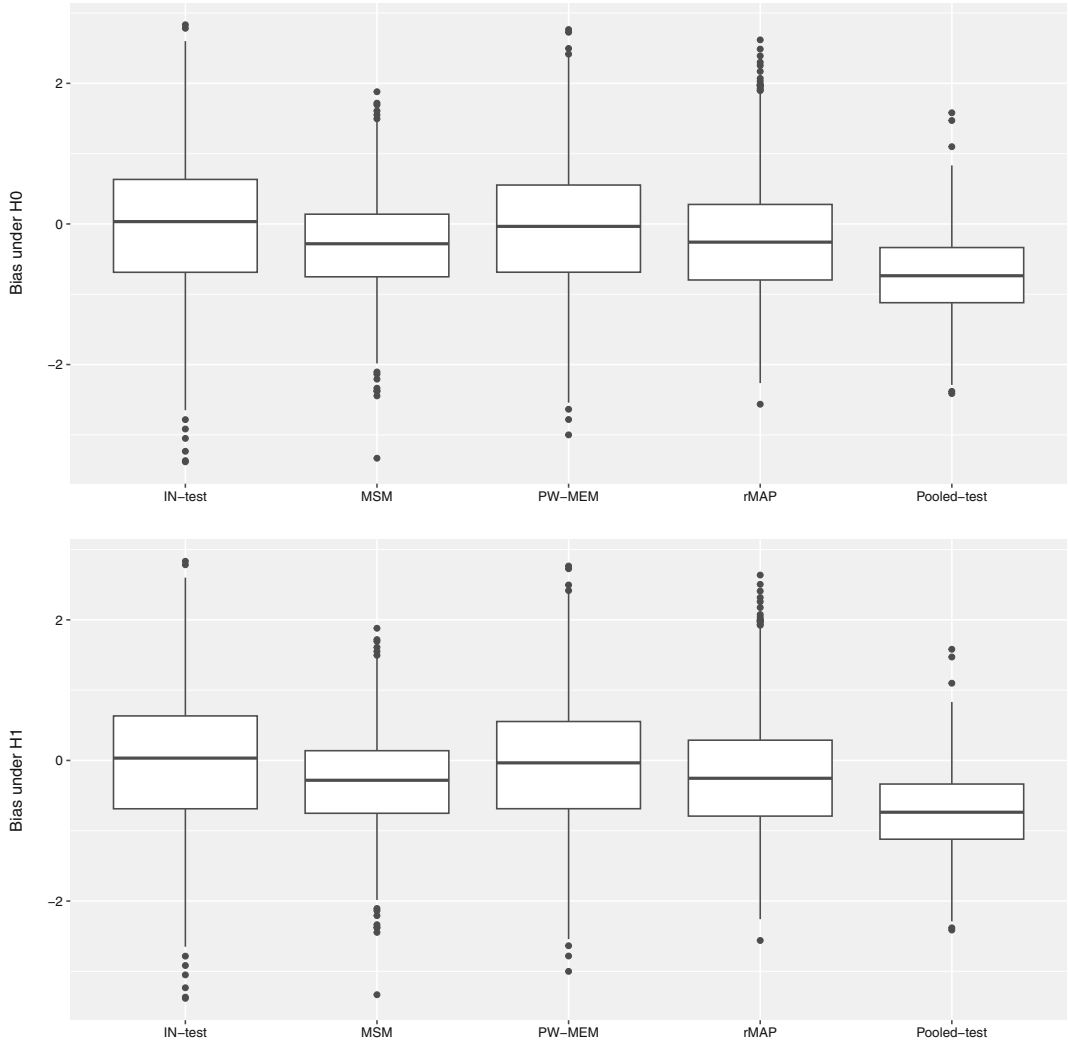

FIGURE 4 The bias of treatment effect estimates using different statistical approaches under the null and the alternative based on bootstrapping from the ADNI dataset.

function becomes less tractable with the inclusion of multiple parameters, and careful evaluations are needed to examine the sensitivity of the marginal likelihood under different prior choices.

A limitation of the proposed strategy is its potential sensitivity to the presence of extreme weights caused by propensity scores near 0 or 1. We recognize that a number of ad hoc decisions can be made to reduce the influence of extreme weights on the construction of the hybrid control arm, including trimming and truncation.[35,36] However, the selection of trimming or truncation threshold is often arbitrary and is therefore out of the scope of this paper. In this work, we demonstrate the proposed PW-MEM approach can be applied to situations when the internal control has a relatively small sample size based on simulation and resampling studies. However, it should be noted that the performance of the PW-MEM method can be negatively affected when the sample size of the internal or external control is too limited. The question of how small is too small requires a case-by-case evaluation. When sample size is too limited, the first step of the PW-MEM approach might produce weighted estimators with increased bias and variance due to the increased chance of empirically violating the positivity assumption regarding to trial participation.[37] The performance of the PW-MEM method also relies on the correct specification of the propensity score model, which should include all the observed confounders and their higher order terms when necessary. Therefore, cautions are needed when implementing the PW-MEM approach to make sure the sample size available can accommodate a propensity score model that provides a reasonably well approximation of the true functional form. In future work, we plan to investigate the sample size requirement for studies employing the PW-MEM method.

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (https://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Wei Wei https://orcid.org/0000-0003-1263-5620
Satrajit Roychoudhury https://orcid.org/0000-0003-4001-3036

REFERENCES

1. US Food and Drug Administration. Rare diseases: natural history studies for drug development guidance for industry. https://www.fda.gov/media/122425/download 2019.
2. Wu J, Wang C, Toh S, Pisa FE, Bauer L. Use of real-world evidence in regulatory decisions for rare diseases in the United States—current status and future directions. *Pharmacoepidemiol Drug Saf*. 2020;29(10):1213-1218.
3. Beckman RA, Natanegara F, Singh P, et al. Advancing innovative clinical trials to efficiently deliver medicines to patients. *Nat Rev Drug Discov*. 2022;21(8):543-544.
4. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Jama*. 2018;320(9):867-868.
5. US Food and Drug Administration. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. https://www.fda.gov/media/164960/download
6. US Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. https://www.fda.gov/media/152503/download
7. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29(3):175-188.
8. Lim J, Walley R, Yuan J, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Ther Innov Regul Sci*. 2018;52:546-559.
9. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15:46-60.
10. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environ: J Int Environ Soc*. 2006;17(1):95-106.
11. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67(3):1047-1056.
12. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7(1):5-18.
13. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70(4):1023-1032.
14. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm Stat*. 2016;15(2):123-134.
15. Ventz S, Comment L, Louv B, et al. The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro Oncol*. 2022;24(2):247-256.

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

16. Ventz S, Khozin S, Louv B, et al. The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nat Commun*. 2022;13(1):1-11.
17. Chen WC, Wang C, Li H, et al. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *J Biopharm Stat*. 2020;30(3):508-520.
18. Liu M, Bunn V, Hupf B, Lin J, Lin J. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Stat Med*. 2021;40(22):4794-4808.
19. Zhu AY, Roy D, Zhu Z, Sailer MO. Propensity score stratified MAP prior and posterior inference for incorporating information across multiple potentially heterogeneous data sources. *J Biopharm Stat*. 2024;34(2):190-204.
20. Li C, Ferro A, Mhatre SK, et al. Hybrid-control arm construction using historical trial data for an early-phase, randomized controlled trial in metastatic colorectal cancer. *Commun Med*. 2022;2(1):90.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
22. Vo TT, Porcher R, Chaimani A, Vansteelandt S. A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis. *Res Synth Methods*. 2019;10(4):582-596.
23. Kaizer AM, Koopmeiners JS, Hobbs BP. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics*. 2018;19(2):169-184.
24. Hobbs BP, Landin R. Bayesian basket trial design with exchangeability monitoring. *Stat Med*. 2018;37(25):3557-3572.
25. Kotalik A, Vock DM, Hobbs BP, Koopmeiners JS. A group-sequential randomized trial design utilizing supplemental trial data. *Stat Med*. 2022;41(4):698-718.
26. Boatman JA, Vock DM, Koopmeiners JS. Borrowing from supplemental sources to estimate causal effects from a primary data source. *Stat Med*. 2021;40(24):5115-5130.
27. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-555.
28. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.
29. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403.
30. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation. 13*. West Sussex, England: John Wiley & Sons; 2004.
31. O'Hagan A, Forster JJ. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Vol 2. West Sussex, England: Arnold; 2004.
32. Weber S, Li Y, Seaman J, Kakizume T, Schmidli H. Applying meta-analytic-predictive priors with the R Bayesian evidence synthesis tools. arXiv preprint arXiv:1907.00603. 2019.
33. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am J Psychiatry*. 1984;141(11):1356-1364.
34. Wang C, Li H, Chen WC, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *J Biopharm Stat*. 2019;29(5):731-748.
35. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.
36. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol*. 2010;172(7):843-854.
37. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020;39(14):1999-2014.

How to cite this article: Wei W, Zhang Y, Roychoudhury S, the Alzheimer's Disease Neuroimaging Initiative. Propensity score weighted multi-source exchangeability models for incorporating external control data in randomized clinical trials. *Statistics in Medicine*. 2024;43(20):3815-3829. doi: 10.1002/sim.10158

10970258, 2024, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.10158 by University Of California, Wiley Online Library on [27/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License