Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Matching design for augmenting the control arm of a randomized controlled trial using real-world data

Yingying Liu, Bo Lu, Richard Foster, Yiwei Zhang, Z. John Zhong, Ming-Hui Chen & Peng Sun

To cite this article: Yingying Liu, Bo Lu, Richard Foster, Yiwei Zhang, Z. John Zhong, Ming-Hui Chen & Peng Sun (2022) Matching design for augmenting the control arm of a randomized controlled trial using real-world data, Journal of Biopharmaceutical Statistics, 32:1, 124-140, DOI: 10.1080/10543406.2021.2011900

To link to this article: https://doi.org/10.1080/10543406.2021.2011900

Published online: 29 Jan 2022.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2022, VOL. 32, NO. 1, 124–140
https://doi.org/10.1080/10543406.2021.2011900

Taylor & Francis
Taylor & Francis Group

Matching design for augmenting the control arm of a randomized controlled trial using real-world data

Yingying Liu[a], Bo Lu[b], Richard Foster[c], Yiwei Zhang[d], Z. John Zhong[e], Ming-Hui Chen[f], and Peng Sun[a]

[a]Global Analytics and Data Sciences, Biogen, Cambridge, Massachusetts, USA; [b]Division of Biostatistics, College of Public Health, the Ohio State University, Columbus, Ohio, USA; [c]Global Analytics and Data Sciences, Biogen, Maidenhead Berkshire, UK; [d]Biostatistics, Apellis Pharmaceuticals, Waltham, Massachusetts, USA; [e]Biometrics, REGENXBIO, Rockville, Maryland, USA; [f]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA

ABSTRACT

Randomized clinical trials (RCTs) have often been considered as the gold standard in drug development, but they may not be fully powered due to limited patient population and can even lead to ethical concerns in rare disease studies. In situations like this, real-world data (RWD)/historical data can be utilized to augment or possibly serve as the control arm for the current trial. If a subset of subjects from the RWD/historical trial could be matched to the concurrent control arm subjects and they are deemed comparable following certain criteria, then pooling the matched subjects from the historical control arm and the concurrent control arm can boost the power. In this paper, we propose two matching methods of borrowing historical control data that not only balance key observed baseline covariates but also ensure the comparability of responses between the historical and concurrent controls. Close similarity in response variables among controls reduces Type I error inflation and provides further protection against unmeasured confounding bias, which is a major challenge in using RWD. Simulation studies are conducted to evaluate the empirical performance of the two matching methods in terms of Type I error rate and power, and an illustrative description of a planned study is presented.

ARTICLE HISTORY
Received 27 May 2021
Accepted 19 November 2021

KEYWORDS
Data augmentation; historical control; matching; propensity score; dynamic borrowing

1. Introduction

Randomized clinical trials (RCTs) have been considered as the gold standard in drug development – well-executed randomized trials can reduce population bias and balance on unmeasured confounding factors so that the treatment difference observed in the trial is mainly attributable to the effectiveness of the experimental intervention. Despite this well-known advantage, RCT can be very costly in terms of time and resource, and sometimes can result in feasibility issues and even lead to ethical concerns (Yang and Yu 2021). In the event that a randomized clinical trial is not feasible or not fully powered, especially for treating a serious rare disease with a high mortality rate and a limited disease population, scientists may consider utilizing real-world data (RWD)/historical data to fully replace the concurrent control or augment the control arm for the current trial (Li et al. 2021; Lin et al. 2018). However, improper use of RWD/historical data can undermine the interpretability of the study result, resulting in a biased treatment effect estimate, an inflated Type I error rate, and a potential reduction in power. In the ICH E10 "Choice of Control Group and Related Issues in Clinical Trials" guidance, FDA highlighted that the suitability of designs with external controls is restricted by the inability to control bias and should be limited to situations where "the effect of treatment is dramatic and the usual course of the disease highly predictable" and "the endpoints are objective and the impact of baseline and

CONTACT Yingying Liu Biogen, yingying.liu@biogen.com 300 Binney Street, Cambridge, Massachusetts 02142, USA

treatment variables on the endpoint is well characterized" (2001). Ghadessi et al. (2020) provided a summary of confirmatory clinical trials using historical controls (HCs) in a recent paper entitled "a roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG)" and found that the majority of them have indications in rare diseases and other common applications for HC in the confirmatory settings include medical devices, label expansion, pediatric indications, and small populations.

To minimize the potential biases caused by using historical data in clinical trials, both frequentist and Bayesian approaches have been developed to enhance the comparability of historical and current data. An intuitive frequentist approach to avoid pooling concurrent control with historical data when they appear to be different is "Test-then-pool" (Viele et al. 2014). The similarity of historical and concurrent control data is compared at a significance level of α and pooling only occurs if the null hypothesis of equality is not rejected. Unlike "test-then-pool" that aims to borrow all information from historical control when deemed appropriate, a variety of propensity score (PS) methods based on the propensity scores originally defined by Rosenbaum and Rubin (1983) provide more flexibility in ways that historical data can be utilized. Propensity score matching selects a subset of historical control subjects who share similar baseline characteristics as those subjects in the current trial. Propensity scores can also be used as a stratification factor (stratification by PS) or a weighting variable (inverse probability treatment weighting), and such methods include all propensity-score-evaluable subjects from the historical control. Because of its versatility, the propensity-score-based methods are one of the most popular frequentist approaches used by researchers to balance baseline characteristics when individual patient data (IPD) is available (Li et al. 2020). However, it is also common with real-world and historical data that access to only aggregate data and summary information on the covariate distribution is available. In such an event that IPD is not available for the historical/external data, the robust indirect comparison methods including the matching-adjusted indirect comparisons (MAIC) method proposed by Signorovitch et al. (2010) and the simulated treatment comparison (STC) method by Ishak et al. (2015) could be used to balance on observed confounders between studies.

Bayesian methods are often used in the context of dynamic borrowing, which discounts the historical information based on levels of similarity between the historical control and concurrent data. Power prior is an informative prior combining the noninformative initial prior and the likelihood of the parameters raised to a power $\gamma \in [0, 1]$ given the historical data. The hyperparameter γ quantifies the heterogeneity between the historical data and the current data and is used as a method of down-weighting historical information when the level of similarity is low. Nikolakopoulos et al. (2018) suggested a method of estimating the power prior for the case when only one historical dataset is available that can control Type I error rate and Wang et al. (2019)proposed to use the propensity score methodology to pre-select and stratify patients from real-world data and apply the power prior to each stratum to obtain the posterior distributions for Bayesian inference. In scenarios with multiple imbalanced prognostic factors, the Bayesian hierarchical models have also been proposed to incorporate patient-level baseline covariates to enhance the appropriateness of the exchangeability assumption between the current and historical control data (Han et al. 2017).

Unlike randomized trials, the treatment assignment mechanism is usually not under control in observational studies. The covariate distributions may differ substantially between the treatment and control groups, which may introduce confounding bias in the causal effect estimation. Matching is a popular method in observational studies to ensure the covariate balance between groups, hence reduce confounding bias (Rosenbaum 2010). Comparing to other adjustments in observational studies, the matching design has the following advantages (Lu 2021): 1) It is more robust as it uses a nonparametric way to balance covariate distributions, which does not rely on parametric outcome models. 2) It resembles the randomization design, which is easily interpretable to clinicians and patients. Moreover, such resemblance ensures the statistical inferential procedures following the randomization design can be applied to the matched data, either parametric or nonparametric. 3) It is more objective as the matching process does not involve the outcome and the causal effect estimation is conducted separately after matching.

In the drug development for severe rare diseases, especially for those diseases with a high mortality rate in the pediatric population (for example, spinal muscular atrophy), it would become unethical to randomize subjects to placebo in clinical trials after an efficacious drug is available on the market. For clinical trials to evaluate a new dosing regimen or therapy, the pivotal trial of the approved regimen naturally provides an option to borrow historical data for the purpose of demonstrating superiority. In the new study, subjects may be randomized with a 2:1 or k:1 ratio between the experimental treatment arm and the active control arm using the approved regimen. At the final analysis, a subset of subjects from the RWD/historical trial could be matched to the concurrent control arm subjects and if they are deemed to be comparable in the response of interest, then the concurrent control arm can be augmented by those matched subjects from the historical control arm for a boost in power. In this paper, we propose two novel matching methods of borrowing historical control data that not only balance key observed baseline covariates, but also ensure the comparability of responses between the historical and concurrent controls, hence provide some protection against unmeasured confounders.

The rest of the paper is organized as follows. In Section 2, we review the concept of propensity score and matching design, introduce an algorithm for matching between a larger target group and a smaller reference group and two matching methods to augment the control arm. In Section 3, we conduct simulation studies and evaluate empirical performances of the two matching methods in different scenarios, in terms of Type I error rate and power. An illustrative description of a planned study is presented in Section 4 and Section 5 is devoted to discussions.

2. Method

2.1. Notation

The randomized population in the current study is indicated as RP with TA denoting the treatment/experimental arm and CA denoting the control arm. For the current study, let Y_R denote the vector of primary outcomes and W_R denote the vector of intermediate/auxiliary outcomes, and the correlation between the outcome variables Y and W is ρ. Let $\mathbf{X_R}$ be the $n_R \times p$ matrix of observed baseline covariates where n_R is the total number of subjects in the current study. Treatment assignment is denoted as T: $T = 0$ indicates the concurrent control arm and $T = 1$ represents the experimental arm in the current study. Let r be the proportion of subjects randomized in the experimental arm and $1 - r$ be the proportion in the concurrent control arm, e.g., $r = 0.5$ for 1:1 randomization and $r = 2/3$ for 2:1 randomization.

For the historical data, let Y_H denote the vector of primary outcomes, W_H denote the vector of intermediate/auxiliary outcomes and $\mathbf{X}_H$ the $n_H \times p$ matrix of observed baseline covariates with n_H indicating the total number of subjects from the external data who satisfy the key inclusion and exclusion criteria of the current randomized study, i.e., the historical control population (HCP). We assume that there is only one external data source and the same p baseline covariates are collected in both the current study and the external study.

Propensity score $\boldsymbol{e}(\boldsymbol{X})$ is a balancing score such that conditional on the propensity score, the baseline covariates in $\boldsymbol{e}(\boldsymbol{X})$ are expected to be balanced between the two treatment groups.

Let i denote the i^{th} subject.

Let $Z_i = 1$ if patient i is from the current study and 0 if patient i comes from the external data.

Let $\boldsymbol{X_i}$ denote the vector of observed baseline covariates for subject i.

Propensity score is defined as the conditional probability of being assigned to treatment group 1 given the observed set of covariates $\boldsymbol{X_i}$, denoted as $\boldsymbol{e}(\boldsymbol{X_i}) = P(Z_i = 1|\boldsymbol{X_i})$. Rosenbaum and Rubin (1983) proved that the treatment assignment and the observed covariates are conditionally independent given the propensity score, i.e., $\boldsymbol{X} \perp Z|\boldsymbol{e}(\boldsymbol{X})$ and that if the treatment assignment is strongly ignorable, then adjustment for $\boldsymbol{e}(\boldsymbol{X})$ or a balancing score $\boldsymbol{b}(\boldsymbol{X})$ finer than $\boldsymbol{e}(\boldsymbol{X})$ is sufficient to produce an unbiased estimate of the average treatment effect. The propensity scores are often

estimated through logistic regression using Z as the dependent variable and $\boldsymbol{X}$ as the independent variables and the predicted probability $P(Z_i = 1|\boldsymbol{X_i})$: $\widehat{\boldsymbol{e}}(\boldsymbol{X_i}) = 1/\left(1 + \exp\left\{-\boldsymbol{X_i}\widehat{\beta}\right\}\right)$ is the estimated propensity score for the i^{th} subject, where $\widehat{\beta}$ is the estimate from the regression model.

2.2. Matching design

Matching refers to a class of methods that group subjects together according to a certain homogeneity criterion with the goal to balance the covariates distribution. Broadly speaking, matching can be done based on the values of propensity score, or any distance metric of covariates $(\boldsymbol{X})$. For example, the degree of homogeneity H_{ij} between any two subjects (i, j) can be defined using distance metric such as (Stuart 2010)

- Euclidean distance: $H_{ij} = \left(\boldsymbol{X_i} - \boldsymbol{X_j}\right)'\left(\boldsymbol{X_i} - \boldsymbol{X_j}\right)$
- Mahalanobis distance: $H_{ij} = \left(\boldsymbol{X_i} - \boldsymbol{X_j}\right)'\Sigma^{-1}\left(\boldsymbol{X_i} - \boldsymbol{X_j}\right)$
- Propensity score distance: $H_{ij} = \left(\boldsymbol{e}(\boldsymbol{X_i}) - \boldsymbol{e}(\boldsymbol{X_j})\right)^2$ or $H_{ij} = \left|\boldsymbol{e}(\boldsymbol{X_i}) - \boldsymbol{e}(\boldsymbol{X_j})\right|$
- Linear Propensity score distance: $H_{ij} = \left(\text{logit}(\boldsymbol{e_i}) - \text{logit}(\boldsymbol{e_j})\right)^2$ or $H_{ij} = \left|\text{logit}(\boldsymbol{e_i}) - \text{logit}(\boldsymbol{e_j})\right|$

The Mahalanobis distance was originally developed for multivariate normal distribution with covariance matrix Σ. However, when certain covariate contains extreme outliers or has a long-tailed distribution, the Mahalanobis distance will tend to ignore that covariate in matching due to its inflated standard deviation. So robust/rank-based Mahalanobis distance was proposed as an alternative measure by replacing the covariates with their ranks and using an adjusted covariance matrix (Rosenbaum 2010).

In observational studies or clinical trials with a historical data borrowing design, researchers often match external datasets (the reference group) of larger sizes with the current study (the target group) with a smaller sample size. There are different matching designs that can be used to match the reference group to the target group (Rosenbaum 2010).

- One-to-1 matching (pair matching) is the most commonly used design. In each matched set, there is exactly one subject from the target group and one from the reference group. This design has good matching quality but may not be very efficient in using the data for large reference groups.
- One-to-k matching can be more efficient in using the external data. In each matched set, there is one subject from the target group and a fixed number of $k > 1$ subjects from the reference group. However, the matching quality may be sacrificed if k is not appropriately selected.
- Variable matching is regarded as an intermediate step between pair matching and 1-to-k matching. In each matched set, there is one subject from the target group and multiple but not fixed number of subjects from the reference group. This matching design use the external data efficiently, but each subject in the target group will carry different weights during testing, which could lead to concerns in some clinical trial settings.
- Full matching makes use of every subject. In each matched set, there is one subject from one group and multiple but not fixed number of subjects from another group. The full matching design uses all available data without sacrificing the matching quality much and it can be viewed as the finest way of stratification with at least one subject from each group.

2.3. An algorithm for matching with a larger target group and a smaller reference group

Generally, matching is conducted between a small target group (typically the treatment group) and a large reference group (typically the control group). This is referred to as matched sampling (Rosenbaum and Rubin 1985), since it is convenient to select a portion of subjects from the larger

reference group to mimic the covariate distribution of a smaller target group. In the rare disease setting, researchers might encounter challenges when the identified reference group (i.e., historical control population) has a smaller sample size than the target group (i.e., current trial participants). Therefore, existing matching algorithms cannot be applied directly. We propose an algorithm to suit the need of matching the historical control population (HCP) of size n_H with the randomized population (RP) in the current study of size n_R where $n_H < n_R$. The basic idea is to first draw a small random sample from RP, then match HCP with this small sample using the conventional matching algorithms to ensure that the distribution of the matched HCP subset mimics the distribution of RP. The process will be repeated a few times and the matched subset resembling RP the most is selected. Let HC (Historical Control) represent the subset of HCP that is borrowed through the matching with sample size n_{HC}. The detailed algorithm is described below.

(1) Specify the number of subjects $n_{HC} < n_H$ to be borrowed from HCP;
(2) Draw random samples of n_{HC} subjects from RP for k times;
(3) Construct matched historical control candidate samples by matching HCP with each random sample based on the key covariates;
(4) Compare the distances between RP and the k matched historical control candidate samples using robust Mahalanobis distance (Pimentel 2016), and select the best match as the HC (the one with the smallest distance).

The more subsamples from RP we generate, the more likely we get a matched HC resembling RP. Although motivated by a rare disease application where the reference group size n_H is smaller than the target group size n_R, this algorithm can be easily applied to borrow an arbitrary number of subjects from the historical control pool which is not proportional to n_R.

2.4. Using matching to augment the control arm from external data

2.4.1. Conditional borrowing

Historical data has rarely been considered as reliable as the current randomized control, even if it was generated from a historical trial done by the same organization as the current study, with largely the same investigators and no apparent dissimilarity in other aspects of the study design (Han et al. 2017). To minimize the concern for some unpredictable bias introduced by the historical data, conditional borrowing can be used to passively control for unmeasured confounding. It is achieved by comparing the primary response variable in the matched control samples based on a pre-specified closeness criterion. If the responses differ substantially (after matching on the key baseline covariates), it is likely due to some unmeasured confounders that are not controlled through matching. The conditional borrowing approach is implemented in the following two steps:

(1) Match the historical data (HCP) with the current trial data (RP) based on the key baseline covariates;
(2) Check the comparability of the matched historical control (HC) and the concurrent control arm (CC) according to a pre-specified closeness criterion for the response variable.

If HC and CC are comparable, then pool subjects from the matched HC with CC and construct the augmented control arm (CA = HC+CC). Otherwise, if HC and CC are deemed different, then borrowing historical data will very likely introduce biases, so the control arm will contain only the concurrent control without pooling any data from the historical control. Figure 1 helps illustrate the conditional borrowing steps discribed above. A key step in the implementation is that the matching process only involves baseline covariates. The response variable is involved after matching as a measure to determine whether pooling is acceptable or not. The specific criterion on determining closeness should be agreed upon among all stakeholders prior to conducting matching. For larger

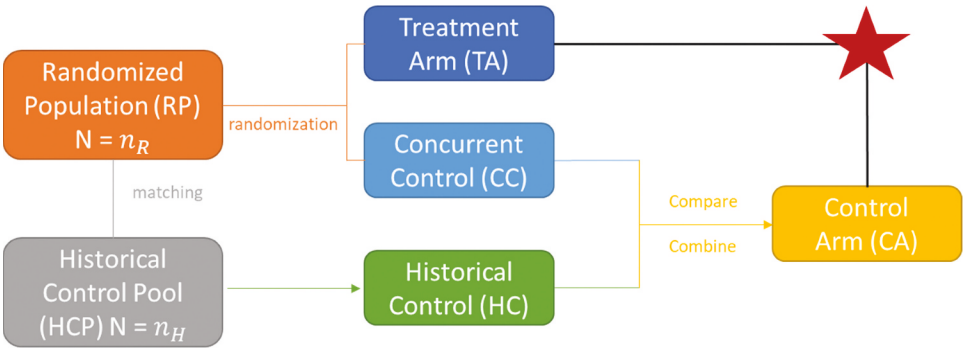

Figure 1. Illustration of conditional borrowing.

concurrent trials where it is safe to assume that the baseline covariates between the treatment arm (TA) and the concurrent control arm (CC) are well-balanced through randomization, propensity score matching in step one of the conditional borrowing approach could also done between HCP and CC .

2.4.2. *Intermediate outcome assisted borrowing*

There could be cases where the decisions on borrowing – the subset of patients to construct the historical control arm and/or whether borrowing would be feasible at final – need to be made before the primary endpoint is available. In such situations, one way to control for unmeasured biases without the primary endpoint is to use some intermediate outcome as one of the matching variables. The intermediate outcomes could be the primary endpoint assessed at an earlier time point or other biomarker or surrogate endpoint that has a moderate/high correlation with the primary endpoint and reflects the treatment effect faster. This approach takes advantage of the early movement of these intermediate outcome measures related to clinical benefit, actively controls for any unmeasured confounding, and allows for accurate decision-making on borrowing before the primary endpoint is available. One important note is that matching only occurs between historical and concurrent control groups and no treatment group data are used, since the intermediate outcome is post-treatment. One needs to carefully check the covariate balance after matching, especially the balance of the intermediate outcome to ensure the comparability between the historical control and concurrent control arms. It would also be helpful to pre-specify an acceptable balance criterion to ensure subjectivity.

2.5. *Post-matching balance checking*

After two groups have been matched, it is important to check the covariate balance. The two commonly used statistical measures are the absolute standardized difference for mean (SDM) and the log-ratio of standard deviations. For continuous and binary variables, the absolute standardized differences for mean are defined as

$$100\left|\bar{X}_{(Z=1)} - \bar{X}_{(Z=0)}\right| / \sqrt{\frac{s^2_{(Z=1)} + s^2_{(Z=0)}}{2}} \tag{1}$$

$$100\left|\hat{p}_{(Z=1)} - \hat{p}_{(Z=0)}\right| / \sqrt{\frac{\hat{p}_{(Z=1)}\left(1-\hat{p}_{(Z=1)}\right) + \hat{p}_{(Z=0)}\left(1-\hat{p}_{(Z=0)}\right)}{2}} \tag{2}$$

where the means in the numerator are calculated from the matched samples and the variances in the denominator are calculated from the pre-matching samples (Rosenbaum and Rubin 1985). The absolute standardized difference should be less than or equal to 0.25 for good variable balance (Stuart 2010) and a smaller threshold of 0.1 has also been used to impose a better balance (Austin 2009; Mamdani et al. 2005; Normand et al. 2001). The log ratio of standard deviations is defined as $\log\left(s_{Z=1}/s_{Z=0}\right)$ and a threshold of 0.2 may be used as it implies that the corresponding variances are within 50% (Rubin 2001).

3. Simulation study

3.1. Simulation setting

In this section, we conduct simulation studies to evaluate the empirical performances of the two proposed matching methods: 1) conditional borrowing and 2) intermediate outcome assisted borrowing. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ denote the key baseline covariates, where $\boldsymbol{x}_1, \boldsymbol{x}_2$ follow independent binary distributions and $\boldsymbol{x}_3, \boldsymbol{x}_4$ follow independent normal distributions. The binary indicator of the treatment status is denoted as T, the primary outcome is denoted as Y and the intermediate/auxiliary outcome is denoted as W.

The joint distribution of (Y, W) follows a bivariate normal distribution with correlation ρ:

$$\begin{pmatrix} \mathrm{Y} \\ \mathrm{W} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\mathrm{Y}} \\ \mu_{\mathrm{W}} \end{pmatrix}, \begin{bmatrix} \sigma_{\mathrm{Y}}^2 & \rho \cdot \sigma_{\mathrm{Y}}\sigma_{\mathrm{W}} \\ \rho \cdot \sigma_{\mathrm{Y}}\sigma_{\mathrm{W}} & \sigma_{\mathrm{W}}^2 \end{bmatrix} \right)$$

It is assumed that the correlation between the primary and intermediate/auxiliary outcomes are the same for the historical data and the current trial.

One historical control population (HCP) with 200 subjects will be simulated. The randomized population (RP) contains 120 subjects randomized with 2:1 ratio to the treatment arm (TA) with $n_{\mathrm{TA}} = 80$ and the concurrent control arm (CC) with $n_{\mathrm{CC}} = 40$. The goal is to identify a historical control (HC) of 40 subjects through matching to potentially augment the control arm for final analysis. For simplicity and the purpose of fair comparison, the historical control pool (HCP) will be matched with the concurrent control (CC) through pair matching. The population indicator Z = 1 represents the randomized population RP in the current study and Z = 0 represents the historical control population HCP. For each choice of $a_0(\mathrm{Z} = 1)$, 1000 RPs will be simulated each under H_0 with $a_5 = 0$ and H_a with $a_5 = 1.5$, so that the power without borrowing historical data is 72.5%.

$$\begin{aligned}
\mu_{\mathrm{Y}}(\mathrm{Z} = 0) &= a_0(\mathrm{Z} = 0) + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 \\
\mu_{\mathrm{W}}(\mathrm{Z} = 0) &= b_0(\mathrm{Z} = 0) + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\
\mu_{Y}(Z = 1) &= a_0(\mathrm{Z} = 1) + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5\mathrm{T} \\
\mu_{\mathrm{W}}(\mathrm{Z} = 1) &= b_0(\mathrm{Z} = 1) + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5\mathrm{T}
\end{aligned}$$

It is also assumed that the baseline covariates have the same effect on the response variables of the two populations, with the intercepts $a_0(\mathrm{Z})$ and $b_0(\mathrm{Z})$ representing any potential unmeasured confounding effect on the responses.

In the simulation, $a_0(\mathrm{Z} = 0)$ is fixed to be 2 and $a_0(\mathrm{Z} = 1) = a_0(\mathrm{Z} = 0) + c$, where c takes values from −0.8 to 4.8 and represents the unknown distributional shift in mean response of the current trial from the historical control under the same control regimen. The correlation ρ between Y and W is set to be 0.2, 0.5 and 0.8 to represent low, medium, and high correlation. Three borrowing strategies were evaluated:

(1) Pooling – match HCP with CC using $x_1 \sim x_4$ and pool matched HC with CC if the absolute SDM of the propensity score distance is less than or equal to 0.1.

(2) Conditional borrowing – match HCP with CC using $x_1 \sim x_4$ and check the absolute SDM of the propensity score distance and the response similarity between the matched HC and CC. Pool HC with CC only when the absolute SDM of the propensity score distance is less than or equal to 0.1 and they are similar in response. The response similarity criterion is set up as

$$\bar{Y}_{CC} \in [\bar{Y}_{HC} - L \times SE(\bar{Y}_{HC}), \bar{Y}_{HC} + L \times SE(\bar{Y}_{HC})] \quad (3)$$

where L takes the values of 1, 1.5 and 2, indicating the number of standard errors the mean of concurrent control arm needs to be from the historical control mean.

(3) Intermediate outcome assisted borrowing – match HCP with CC using $x_1 \sim x_4$, W and check the balance of W and the propensity score distance using the standardized difference for mean. Pool HC and CC only if the absolute SDMs of W and the propensity score distance are both less than or equal to 0.1.

3.2. Simulation results

3.2.1. Pooling and conditional borrowing

Figure 2 presents the two-sided Type I error profiles of borrowing historical control data using pooling and conditional borrowing. The pooling method corresponds to an extreme case of conditional borrowing where L = ∞ (literally no restriction on mean similarity), denoted as "inf" in Figure 2. The x-axis shows the population mean response of the concurrent control arm (CC) in the Type

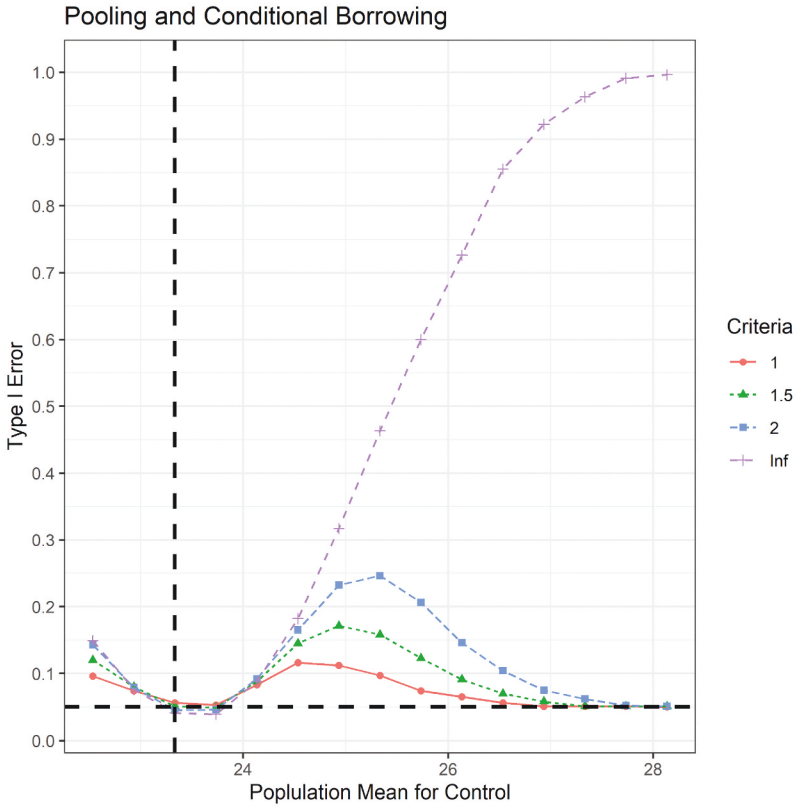

Figure 2. Type I error rate for pooling and conditional borrowing.

I error plots, with the vertical dotted line indicating the mean response of the historical control population (HCP). To illustrate the overall Type I error profile, the simulation results cover a wider range of the current trial mean response than what is plausible to be observed in reality.

As shown in Figure 2, with both methods, Type I error rate is controlled at the two-sided 0.05 significance level when the CC mean is the same as the HCP mean and starts to inflate when the CC mean shifts away from the HCP mean. Without mean similarity check, the Type I error rate of the pooling method goes up to one when there is a substantial difference in the mean response between the current trial and the historical data resulted from some unmeasured confounding variable. However, with the additional mean similarity check on the response variable, for conditional borrowing, the Type I error rate is bounded. The stringency of the L-criterion determines both the magnitude of Type I error inflation and the position where the maximum Type I error rate occurs – the amount of inflation resulted from a looser criterion of $L = 2$ doubles the amount from a strict criterion of $L = 1$. Moreover, with a more stringent criterion, the maximum Type I error rate is reached with a smaller CC mean deviation from the HCP mean. With a two-sided test, the Type I error rate is also inflated when the CC mean is less than the HCP mean, but it can be avoided with a one-sided test.

In addition to the Type I error rates, Figures 3 and 4 present the borrowing rate (empirical probability of borrowing), power, and treatment effect estimation % bias defined as $(\hat{a}_5 - a_5)/a_5$ of the conditional borrowing method with different standard error thresholds. The borrowing rate is the highest when the CC mean equals to the HCP mean, resulting in a boost in power of almost 10% even with the strict criterion of $L = 1$, and the empirical probability of borrowing drops down to zero when the CC mean deviates from the HCP mean. These results suggest that the conditional borrowing method provides a decent amount of power increase with no Type I error inflation when the distribution of the current control data is the same as the historical data, and it is capable of detecting the overall distributional shift of CC from HCP and adjust the probability of borrowing accordingly. Due to the relatively high borrowing rate when the CC mean is slightly higher than the HCP mean, further power increase can be obtained at the cost of minimal Type I error inflation. Even if the CC mean is slightly lower than the HCP mean, borrowing from historical data still provides some power gain because the loss in mean difference can be offset by the gain in smaller pooled SD with a high borrowing rate. However, if the CC mean is much lower than the HCP mean, borrowing from the historical data would result in power loss. The estimation bias curves are similar in shape as the Type I error curves with the maximum bias achieved when the CC mean is moderately different from the HCP mean and there is still over 40% of borrowing from the HC.

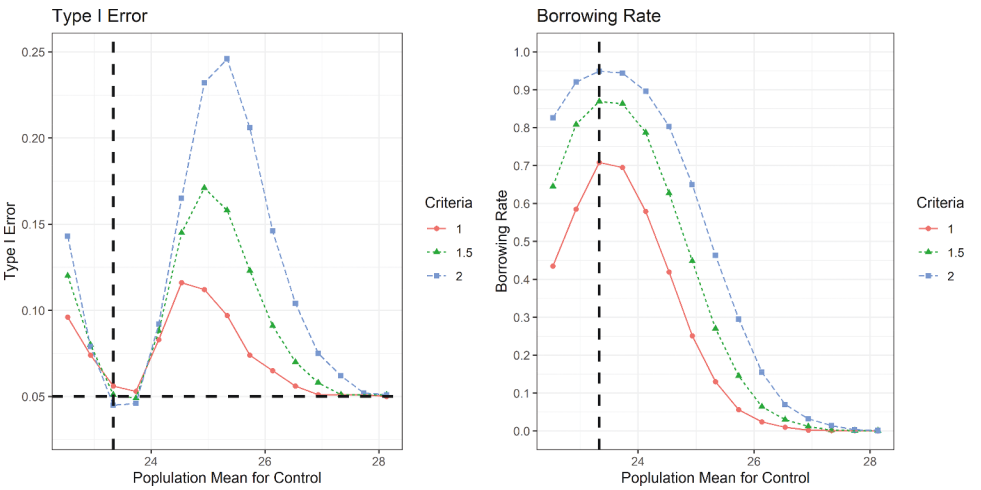

Figure 3. Type I error rate and borrowing rate for conditional borrowing.

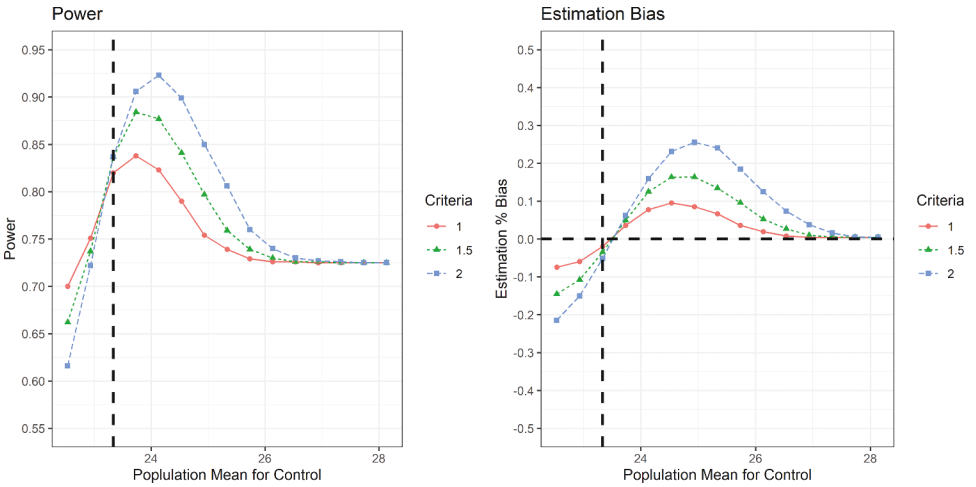

Figure 4. Power and estimation bias for conditional borrowing.

3.2.2. Intermediate outcome assisted borrowing

Figure 5 shows the Type I error inflation and borrowing rate using the intermediate outcome assisted borrowing method with the absolute standardized mean difference (SDM) threshold of 0.1 for post matching balance check. With the absence of primary endpoint, using an intermediate outcome as one of the propensity score matching variables and assuring good balance between HC and CC results in bounded Type I error inflation. However, the maximum inflation is highly dependent on the strength of correlation between the intermediate and primary outcome, and the Type I error rate is only under reasonable control when the correlation is high. Based on the borrowing rate curves on the right panel, a higher correlation means better sensitivity to the overall distributional shift of CC from HCP and an earlier drop in the probability of borrowing when the CC mean shifts away from the HCP mean.

Figure 6 demonstrates the power and treatment effect estimation bias for the intermediate outcome assisted borrowing method with different correlation assumptions. The power curve under the higher correlation scenario of 0.8 is similar in shape as those of the conditional borrowing method above, with a decent power boost of 7.5% when the CC mean is the same as

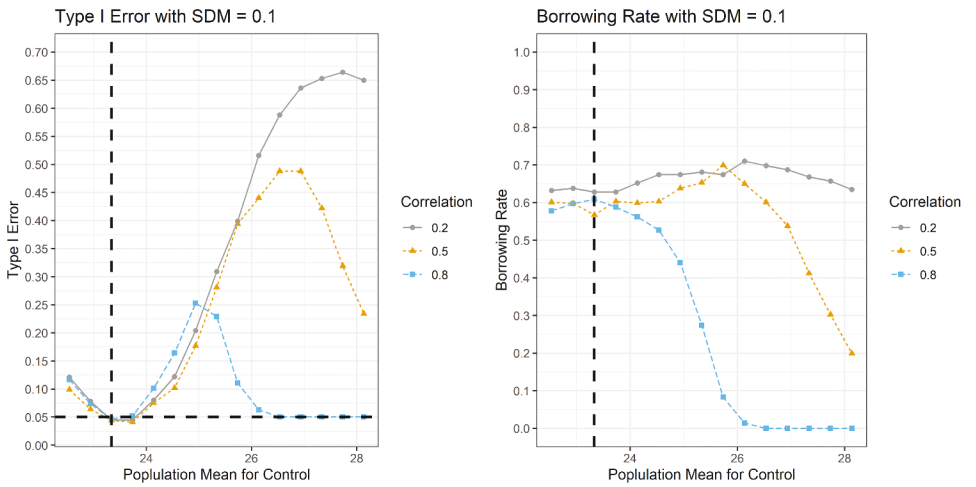

Figure 5. Type I error rate and borrowing rate for intermediate outcome assisted borrowing with SDM threshold of 0.1.

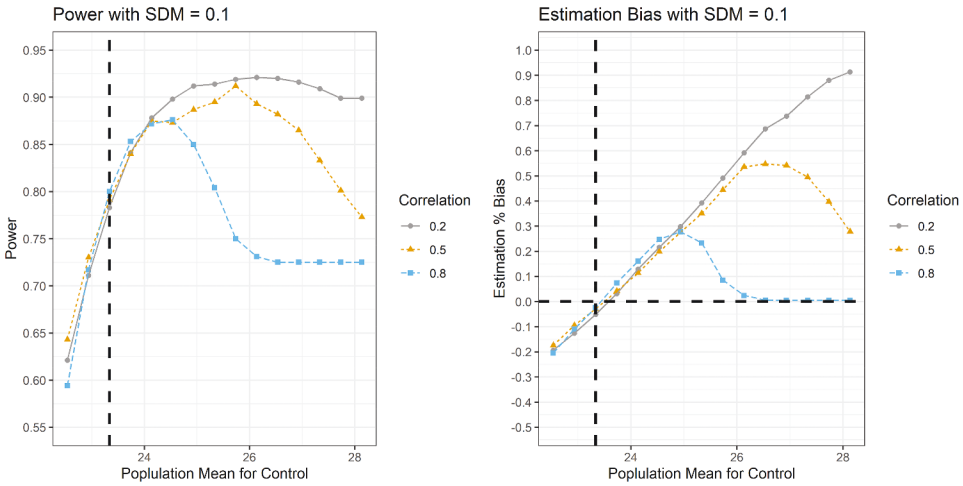

Figure 6. Power and estimation for intermediate outcome assisted borrowing with SDM threshold of 0.1.

the HCP mean. Thus, if the primary endpoint is not available at the time when the decision on borrowing needs to be made and an intermediate endpoint highly correlated with the primary endpoint is available, this outcome assisted borrowing method could shed some light on the similarity between the historical data and the concurrent control data and result in a more informed decision.

3.2.3. *Matching a larger target group with a smaller reference group*

To demonstrate the performance of the proposed method for matching with a larger target group and a smaller reference group, a second historical control population (HCP) with $n_H = 60$ subjects is simulated. The 1000 simulated randomized populations (RP) with $n_R = 120$ subjects remains the same as those in the previous simulation setting. The goal is to identify a historical control (HC) of size $n_{HC} = 40$ through matching HCP with RP using covariates $x_1 \sim x_4$. The number of random draws (k) from the randomized population is set to be 10 and 50 respectively, and similar conditional borrowing method as demonstrated in Section 3.1 strategy (2) is used – pool HC with CC only when the absolute SDM of the propensity score distance is less than or equal to 0.1 and they are similar in response as defined in Equation 3. Figure 7 presents the Type I error rate, probability of borrowing and power for this method with different choices of k.

It can be observed that the overall shapes of all the curves have the same pattern as those presented in Section 3.2.1 – Type I error inflation is bounded, and the power boost is meaningful when the population mean of the concurrent control arm is the same as that of the historical control pool. Due to the fact that the historical control pool is small, with the restriction on the absolute SDM of the propensity score distance to ensure matching quality, borrowing rate is generally lower. Comparing the left and right panel, one can observe that the performances for $k = 10$ is very similar to that for $k = 50$. Thus, it is safe to conclude that in small sample size setting, if the goal is to select a relatively big historical control arm from the historical control pool, it does not require a large number of random draws. Although motivated by a rare disease application where the reference group size n_H is smaller than the target group size n_R, this algorithm can be easily applied to borrow an arbitrary number of subjects from the historical control pool, which is not proportional to n_R.

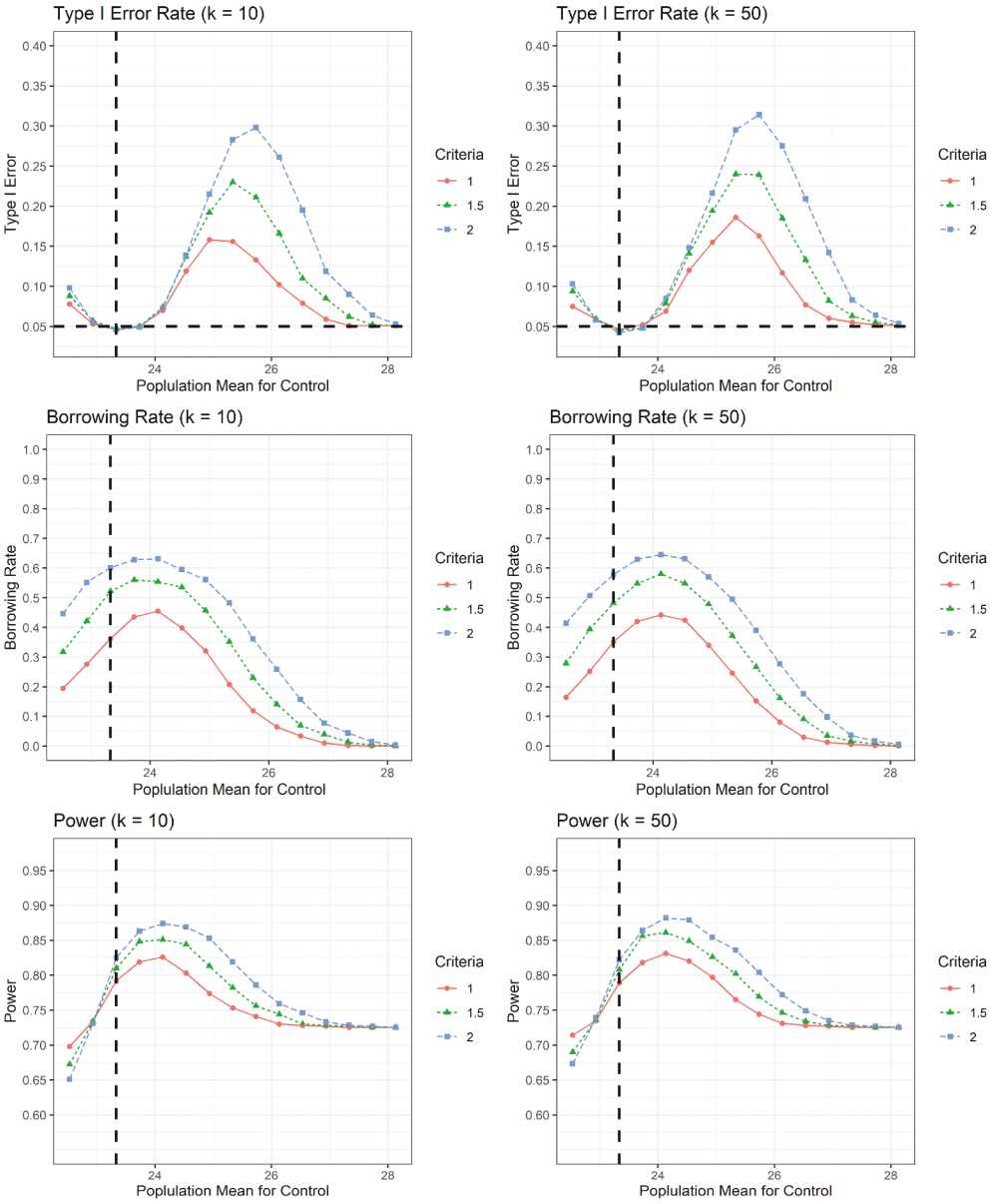

Figure 7. Type I error rate, borrowing rate and power for k = 10 and 50.

4. An illustrative description of a planned study

To illustrate the implementation of our proposed methodology, we consider a randomized study of relapsed systemic anaplastic large cell lymphomas (sALCL) with a design feature to potentially borrow from RWD. Anaplastic large cell lymphoma (ALCL) is a very rare disease that accounts for approximately 3% of the cases of adult non-Hodgkin lymphoma (NHL). Unlike standard ALCL, sALCL is not localized and thus required systemic therapy. Due to the rare lymphomas type and ethical considerations, it is difficult to conduct a large randomized phase 3 clinical trial. After clinical, statistical, and trial feasibility evaluations, a smaller randomized study that has the potential to borrow RWD to augment the control arm is considered.

The primary endpoint is progression-free-survival (PFS), which is defined as the time from the start of study treatment to the first documentation of objective tumor progression or to death due to any cause, whichever comes first. The secondary endpoint is objective response rate (ORR), which is defined as the proportion of patients with complete remission (CR) or partial remission (PR) according to the International Working Group (IWG) Revised Response Criteria for Malignant Lymphoma (Cheson 2007). At the design stage, appropriate RWD sources and seven important baseline covariates are identified through literature review, which includes age, gender, race, anaplastic lymphoma kinase (ALK) status, baseline B Symptoms, ECOG performance status (a scale developed by the Eastern Cooperative Oncology Group), and prior line of treatment. Assuming an exponential distribution of PFS, 115 events will be required to detect a hazard ratio of 0.6 with approximately 80% power using the two-sided α level of 0.05. However, due to the limited patient population and slow recruitment rate, only 90 subjects are expected to be recruited in this randomized clinical trial, which is less than the event size needed to achieve sufficient power. Therefore, the study team decide to randomize 90 patients in 2:1 (treated: SOC) fashion, and potentially borrow additional 30 patients from historical data to augment the control group and increase the overall power. Based on the similarly of inclusion and exclusion criteria, prior treatment, cancer types and endpoint availability, a total of 100 patients are identified from the RWD database.

For conditional borrowing, the median progression-free survival time were used to check response similarity: if the median progression-free survival time of the concurrent control is within the 68% confidence interval of the median progression-free survival time of the historical control after matching, then it is deemed feasible to borrow. Note that this is equivalent to the similarity criterion of $L = 1$ in the continuous response variable case. Since ORR is directly attributable to drug effect, it has been the most commonly used surrogate endpoint in support of accelerated approval in oncology clinical trials (FDA 2018). Based on the fact that ORR is highly correlated with the primary survival endpoint and can be observed earlier in time, it is used as the intermediate outcome W for the outcome assisted borrowing.

At the final analysis, the effect of the treatment comparing to the SOC was estimated by the hazard ratio (HR). Figure 8 is the Kaplan–Meier curves for the current study of 90 subjects and the historical control pool of 100 subjects. Without borrowing from the historical control pool, the hazard ratio is 0.65 with a p-value of 0.061. However, if conditional borrowing method is used, the median progression-free survival time of the two control arms satisfy the pre-specified similarity criteria and the resulted HR between the treatment arm and the augmented control arm of SOC is 0.69 (p-value = 0.048). If the intermediate outcome assisted borrowing method is applied, the post-

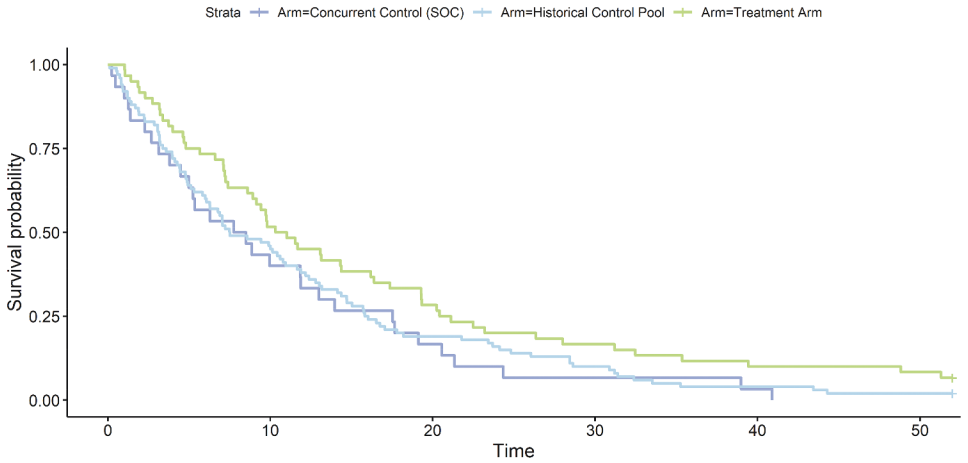

Figure 8. Kaplan meier curves for the current study (Treatment am and concurrent control) and the historical control pool.

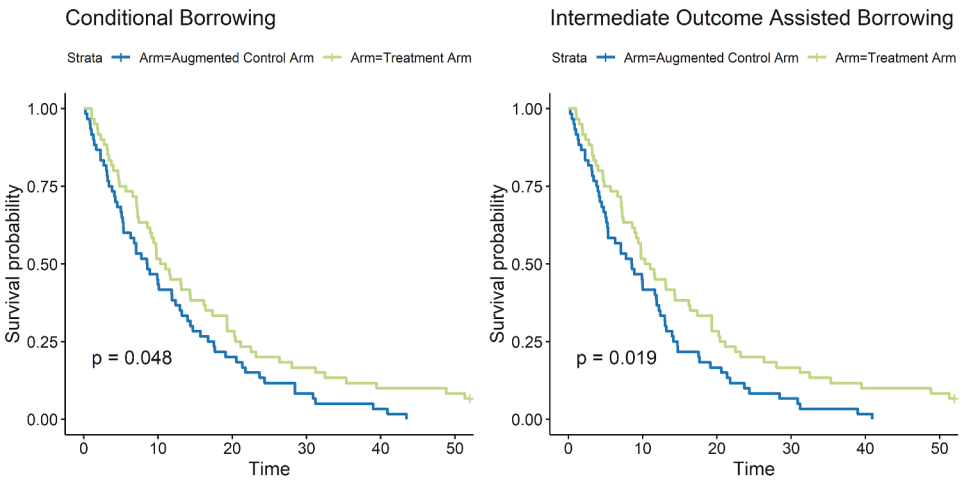

Figure 9. Kaplan meier curves after borrowing from historical data.

matching absolute SDMs of the ORR and the overall propensity score distance are both less than 0.1. In fact, further similarity check on the primary time-to-progression endpoint shows that the median progression-free survival time of the concurrent control is within the 68% C.I. of the historical control, thus satisfying the conditional borrowing criteria as well. The resulted HR between the treatment arm and the augmented control arm of SOC is 0.64 with a significant p-value of 0.019. Figure 9 presents the Kaplan–Meier curves after borrowing from historical data. This example shows that if a study is underpowered, borrowing from historical data that is similar to the concurrent control could help increase the probability of trial success.

5. Discussion

This paper introduced two matching methods of borrowing RWD or historical data that not only balance the key observed baseline covariates, but also ensure the comparability of responses between the historical and concurrent controls. The latter is crucial in RWE. Due to the lack of randomization, unmeasured confounding is a major threat to the validity of clinical findings. The conditional borrowing method is a passive way of protecting against unmeasured confounders when the final efficacy endpoint is available. On the other hand, if decisions on borrowing need to be made at the interim when the primary endpoint is not yet available, the intermediate outcome assisted borrowing can be applied to actively control for unmeasured confounding, given there exists a potential surrogate or intermediate endpoint that is correlated with the primary endpoint. The intermediate endpoint can also be a biomarker or the primary efficacy endpoint itself assessed at an earlier time point when the complete data is available, but it is critical to select one that is highly correlated with the primary endpoint.

In the simulation study for outcome assisted borrowing in Section 3.2.2, for illustrative purposes, the trial design is considered such that at the interim, the matched historical control arm and the feasibility to borrow are determined using the intermediate endpoint and the final evaluation of efficacy is fully based on this interim decision. In a real trial design, however, when the primary efficacy endpoint becomes available at the end of the study, one may add a conditional borrowing step to further check for response comparability before pooling the HC matched at the interim. Even if there is no need for an interim decision on borrowing and the conditional borrowing method is used, when the historical control population is large enough and there exists a surrogate endpoint (W) that is known to be highly correlated with the primary endpoint (Y), it is still sensible to include this endpoint (W) as a matching variable. Furthermore, the two proposed methods can be combined in an adaptive design with the option of sample size increase:

- At the interim, use the intermediate outcome as a variable in propensity score matching and determine if the matched historical control data is similar to the concurrent control data through covariate balance check;
- If the post-matching matching balance is reasonably well, then proceed with the original sample size and pool the matched HC with CC for the final analysis;
- Otherwise, if after matching, the intermediate outcome or the overall propensity score distance does not satisfy the pre-specified closeness criteria, then increase sample size to achieve the desired power. At the final analysis, use the conditional borrowing approach and attempt to borrow from historical data with an increased sample size.

For the above adaptive design to be applied, further simulation studies need to be conducted to understand the impact on the Type I error rate.

In recent years, Bayesian methods have gained increasing popularity in dynamic historical data borrowing. Even though they have the ability to adaptively adjust the amount of information borrowed from historical data, the proposed conditional borrowing method has the following advantages: 1) easy to implement 2) flexible enough to be applied to any analytical approach for final analysis and any other endpoints and 3) easy to be communicated with team members and clinicians. For clinical trials that have missing data due to death and need to use some type of composite method to account for both clinical function endpoint and survival, the conditional borrowing method can be easily extended to ensure comparability for two or more response variables. Successful matching creates a randomization-like scenario, and individual level data from matched subjects can be used. Therefore, any analytical methods used for testing the primary endpoint in randomized trials can be applied to the post-matching data. In fact, in the FDA guidance for industry "Amyotrophic Lateral Sclerosis: Developing Drugs for Treatment" (2019), joint rank test (Berry et al. 2013) is specifically required as the single overall measure:

> Sponsors should characterize an effect on mortality in all ALS development programs because it is important to the consideration of the overall safety and effectiveness profiles. If patient function is intended to be assessed by the primary outcome, mortality should be integrated into the primary outcome by an analysis method that combines survival and function into a single overall measure, such as the joint rank test (see section III.B.4.b., Integrated assessment of function and survival).

Moreover, it is straightforward to perform further tests on all other secondary and exploratory endpoints using the augmented data, whether they are continuous, binary or time-to-event. Thus, for drug development especially in the rare disease setting where mortality needs to be characterized, the proposed conditional borrowing approach has the capability to borrow additional RWD/historical data, increase the probability of success for researchers and ethically reduce patient burden.

Although the proposed approaches show the potential to augment the traditional randomized clinical trials with RWD, it is still important to pay additional attention to the practical considerations of using the propensity score methods in the clinical development such as baseline covariates, sensitivity analysis, practical implementation flow, etc (Li et al. 2020).

Disclosure statement

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

Disclaimer

Dr. Z. John Zhong contributed to this project during his employment with Biogen. The views and opinions expressed may not necessarily represent the views of any subsequent employer of Dr. Zhong.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Ming-Hui Chen http://orcid.org/0000-0003-1935-2447

References

Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment Groups in propensity-score matched samples. *Statistics in Medicine* 28 (25):3083–3107. doi:10.1002/sim.3697.

Berry, J. D., R. Miller, D. H. Moore, M. E. Cudkowicz, L. H. van Den Berg, D. A. Kerr, Y. Dong, E. W. Ingersoll, and D. Archibald. 2013. The combined assessment of function and survival (CAFS): A new endpoint for ALS clinical trials. *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 14 (3):162–168. doi:10.3109/21678421.2012.762930.

Cheson, B.D., B. Pfistner, M. E. Juweid, R. D. Gascoyne, L. Specht, S. J. Horning, B. Coiffier, R. I. Fisher, A. Hagenbeek, E. Zucca, R. T. Hoppe, T. A. Lister, M. Dreyling, K. Tobinai, J. M. Vose, J. M. Connors, M. Federico, V. Diehl, V. Diehl, International Harmonization Project on Lymphoma. 2007. Revised response criteria for malignant lymphoma. Journal of Clinical Oncology 25 (5):579–86. doi: 10.1200/JCO.2006.09.2403.

FDA. 2001. *Guidance for industry: E 10 choice of control group and telated issues in clinical trials*. Rockville, Md: U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER).

FDA. 2018. *Clinical trial endpoints for the approval of cancer drugs and biologics*. Silver Spring, MD: U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER).

FDA. 2019. *Amyotrophic lateral sclerosis: Developing drugs for treatment: Guidance for industry*. Silver Spring, MD: U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER).

Ghadessi, M., R. Tang, J. Zhou, R. Liu, C. Wang, K. Toyoizumi, C. Mei, L. Zhang, C. Q. Deng, and R. A. Beckman. 2020. A roadmap to using historical controls in clinical trials - by drug information association adaptive design scientific working group (DIA-ADSWG). *Orphanet Journal of Rare Diseases* 15 (1):69. doi:10.1186/s13023-020-1332-x.

Han, B., J. Zhan, Z. J. Zhong, D. Liu, and S. Lindborg. 2017. Covariate-adjusted borrowing of historical control data in randomized clinical trials. *Pharmaceutical Statistics* 16 (4):296–308. doi:10.1002/pst.1815.

Ishak, K. J., I. Proskorovsky, and A. Benedict. 2015. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* 33 (6):537–549. doi:10.1007/s40273-015-0271-1.

Li, Q., G. Cheng, J. Lin, A. Chi, and S. Davies. 2021. External control using RWE and historical data in clinical development. In *Real-world evidence in drug development and evaluation*, ed. H. Yang, and B. Yu. Boca Raton and London: CRC Press. 71–99 . doi:10.1201/9780429398674.

Li, Q., J. Lin, A. Chi, and S. Davies. 2020. Practical considerations of utilizing propensity score methods in clinical development using real-world and historical data. *Contemporary Clinical Trials* 97:106–123. doi:10.1016/j.cct.2020.106123.

Lin, J., M. Gamalo-Siebers, and R. Tiwari. 2018. Propensity score matched augmented controls in randomized clinical trials: A case study. *Pharmaceutical Statistics* 17 (5):629–647. doi:10.1002/pst.1879.

Lu, B. 2021. Causal inference for observational studies/real-world data. In *Real-world evidence in drug development and evaluation*, ed. H. Yang, and B. Yu. Boca Raton and London: CRC Press 129–148 . doi:10.1201/9780429398674.

Mamdani, M., K. Sykora, P. Li, S. L. Normand, D. L. Streiner, P. C. Austin, P. A. Rochon, and G. M. Anderson. 2005. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ* 330 (7497):960–962. doi:10.1136/bmj.330.7497.960.

Nikolakopoulos, S., I. van der Tweel, and K. C. B. Roes. 2018. Dynamic borrowing through empirical power priors that control type I error. *Biometrics* 74 (3):874–880. doi:10.1111/biom.12835.

Normand, S. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil. 2001. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology* 54 (5):387–398. doi:10.1016/s0895-4356(00)00321-8.

Pimentel, S. D. 2016. Large, sparse optimal matching with R package rcbalance. *Observational Studies* 2:4–23. doi:10.1353/obs.2016.0006.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi:10.1093/biomet/70.1.41.

Rosenbaum, P. R., and D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39 (1):33–38. doi:10.2307/2683903.

Rosenbaum, P. R. 2010. *Design of observational studies*. New York: Springer-Verlag.

Rubin, D. B. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2:169–188. doi:10.1023/A:1020363010465.

Signorovitch, J. E., E. Q. Wu, A. P. Yu, C. M. Gerrits, E. Kantor, Y. Bao, S. R. Gupta, and P. M. Mulani. 2010. Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with Adalimumab or etanercept. *Pharmacoeconomics* 28 (10):935–945. doi:10.2165/11538370-000000000-00000.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25:1–21. doi:10.1214/09-STS313.

Viele, K., S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. 2014. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* 13 (1):41–54. doi:10.1002/pst.1589.

Wang, C., H. Li, W. Chen, N. Lu, R. Tiwari, Y. Xu, and L. Q. Yue. 2019. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics* 29 (5):731–748. doi:10.1080/10543406.2019.1657133.

Yang, H., and B. Yu Eds. 2021. *Real-world evidence in drug development and evaluation*. Boca Raton and London: CRC Press. doi:10.1201/9780429398674.