Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **73**, 444–459
https://doi.org/10.1093/jrsssc/qlad103
Advance access publication 17 November 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Adaptive hybrid control design for comparative clinical trials with historical control data

Beibei Guo[1], Glen Laird[2], Yang Song[2], Josh Chen[2] and Ying Yuan[3]

[1]Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, USA
[2]Vertex Pharmaceuticals, Boston, MA 02210, USA
[3]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Address for correspondence: Ying Yuan, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. Email: yyuan@mdanderson.org

Abstract

We propose an adaptive hybrid control causal (AHCC) design to leverage historical control data to reduce the sample size demanded by standard randomised controlled trials (RCT). Under the causal inference framework, we define the causal estimand of the average treatment effect and derive the corresponding estimator based on the trial data and historical control data. The AHCC design takes a multistage or group sequential approach. The number of patients randomised to the concurrent control is adaptively adjusted based on the amount of information borrowed from the historical control data. At each stage, based on the interim data, the contribution of the historical control data, quantified by the effective sample size, is updated and used to determine the randomisation ratio between the treatment and control arms for the next stage, with the goal to resemble a standard RCT upon the completion of the trial. Simulation studies show that the AHCC design has desirable operating characteristics. For example, it saves on sample size when substantial information can be borrowed from the historical control, and it maintains power when little information can be borrowed from the historical control.

Keywords: balancing weights, effective sample size, historical control, propensity score, randomised controlled trial, real-world evidence

1 Introduction

A randomised controlled trial (RCT) is the gold standard to evaluate and estimate the causal effect of a new treatment compared to a control (e.g. the standard of care). An RCT, however, often requires large sample sizes and is thus often not suitable for rare diseases, defined as a disease or condition that affects less than 200,000 people in the USA per the Orphan Drug Act. This issue also arises in precision medicine when the objective is to evaluate the treatment effect in a specific genomic subtype of a disease. The disease itself may not be rare, but the subtype may be rare.

Leveraging historical data provides an important approach to circumvent this issue. This article focuses on the use of historical control data to reduce the sample size of concurrent controls in RCTs, sometimes known as hybrid control/design, augmented, or synthetic control. This approach has been embraced by industry and regulatory authorities. The US Food and Drug Administration (FDA) has released guidance in the 'Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices' (US Food and Drug Administration, 2017b), and a draft guidance on 'Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry' (US Food and Drug Administration, 2021). Several drugs have been approved based on the synthetic control approach. For example, cerliponase alfa was approved for treating a specific form of Batten disease based on comparison of data from a single-arm trial with 22 patients compared to 42 patients in an

Received: May 20, 2022. **Revised:** August 15, 2023. **Accepted:** October 28, 2023

external control group (US Food and Drug Administration, 2017a). Blinatumomab was approved as a treatment of Philadelphia chromosome-negative relapsed or refractory precursor B-cell acute lymphoblastic leukaemia based on a single-arm trial compared with a synthetic control sample extracted from 13 historical study groups (Przepiorka et al., 2015). Palbociclib was approved by the FDA for the treatment of men with HR + , HER2− metastatic breast cancer using synthetic control data (Pfizer, 2019).

Propensity score matching is one of the most commonly used approaches to incorporate the historical control data in the RCT. Matched pairs of patients, one in the treatment arm and the other from the historical control, may be identified if the difference of their propensity scores is smaller than a predetermined value, called a caliper. Then, the treatment effect can be evaluated by comparing the treatment arm versus the matched historical controls or the combination of matched historical control and concurrent control (known as hybrid control). If the size of a historical control dataset is large enough, it may be possible to find two or more control patients that are close matches for each patient in the treatment arm, so a 2-to-1 or 3-to-1 matching may be done to improve reliability. The related work includes Stuart and Rubin (2008), Yuan et al. (2019), Lin et al. (2019), and Chen et al. (2020), among others. The other major line of methodological development are the Bayesian information-borrowing methods. Ibrahim and Chen (2000) and Ibrahim et al. (2003) proposed a power prior to borrow information from historical control. Hobbs et al. (2011) proposed a commensurate prior that centres the parameter of trial data at that of the historical data, which could be regarded as a special form of the Bayesian hierarchical model. Thall et al. (2003), Berry et al. (2013), and Chu and Yuan (2018a, 2018b) proposed to borrow information from different data resources or subgroups using the Bayesian hierarchical model. Jiang et al. (2023) proposed the elastic prior to achieve dynamic information borrowing by letting the information borrowed depend on the congruence between trial data and historical control. Wang et al. (2019) described the propensity score-integrated power prior approach for incorporating historical data in single-arm clinical studies. Their approach was then extended by Li et al. (2021) for the augmentation of both arms of a RCT and by Lu et al. (2022) for the situation of multiple external data sources. Wang et al. (2022) proposed propensity score-integrated Bayesian prior approaches for augmented control designs. Most aforementioned methods focus on data analysis and statistical inference, as opposed to trial design. In addition, few of them consider the estimand framework recommended by the International Council for Harmonization (ICH) E9 and regulatory agencies.

This article focuses on trial design and aims to address the practical problem of not knowing how much information can be borrowed from the historical control at the design phase of the trial. This is because the trial patient data are not yet available. Thus, it is challenging to choose an appropriate sample size when designing and planning the trial. If we choose a large sample size, we may end up with long enrolment times as well as excessive power if substantial information can be borrowed from the historical control, defeating the goal of using historical data to save on the sample size. On the other hand, if we choose a small sample size, we may end up with an underpowered study if limited information can be borrowed from the historical data. In many cases, only when the trial completes is it recognised that there is too little information that can be borrowed from the historical control to provide an adequately powered comparison of the treatment and control.

We propose an adaptive hybrid control causal (AHCC) design to address this issue, which allows the sample size of the trial to be adaptively adjusted based on the 'usefulness' of the historical control. The objective of AHCC is to emulate a standard RCT (denoted as the target RCT) using a potentially smaller sample size by leveraging the historical control data. Under the causal inference framework, we define the causal estimand of the average treatment effect and derive the corresponding estimator based on the trial data and historical control data. The contribution of the historical control data is quantified using the effective sample size (ESS). The AHCC design takes a multistage (or group sequential) approach. At each stage, based on the interim data, the ESS of historical data is updated and used to determine the randomisation ratio between the treatment and control arms for the next stage, with the goal of resembling the target RCT at the completion of the trial. Simulation studies show that the proposed design has desirable operating characteristics. It saves on sample size when substantial information can be borrowed from the historical control, and it maintains power when little information can be borrowed from the historical

control. The type I error rate is controlled. One difference between our methods and the previous methods that use propensity scores to incorporate historical data is that we use propensity score weighting to map historical data to the current trial data, while others either use propensity score matching or construct an informative prior using historical data.

The remainder of this article is organised as follows. In Section 2, we describe the causal estimand and corresponding estimator of the average treatment effect with historical control data and the AHCC design. In Section 3, we investigate the operating characteristics of the AHCC design through a simulation study. We provide concluding remarks in Section 4.

2 Methods

2.1 Causal estimate of the treatment effect with historical data

Consider a two-arm RCT comparing a new treatment T versus a control C (e.g. the standard of care) in a target patient population $\mathcal{P}$. We assume that patients enrolled in the RCT are a random sample from $\mathcal{P}$. We assume that historical control data were collected from previous studies of C in a patient population $\mathcal{Q}$, which may be the same or partially overlap with $\mathcal{P}$. For example, the RCT and previous studies may use different eligibility criteria for enrolment, or use the same eligibility criteria but patient characteristics drift over time. Let $U = 1$ or 0 denote the current RCT and historical control, respectively, and $Z = 1$ or 0 indicate that a patient received T or C, respectively. Note for all patients in the historical control group, $Z = 0$. Let $\mathbf{x}$ denote baseline covariates and Y denote the outcome of interest. In this article, we consider continuous or binary Y.

Under the potential outcome framework (Imbens & Rubin, 2015; Rubin, 1974, 1978), we assume the existence of potential outcomes $Y_i(0)$ and $Y_i(1)$ for each patient i. These represent the outcome that would be observed under assignment $Z = 0$ or 1, respectively. We make the standard stable unit treatment value assumption (SUTVA; Rubin, 1980) to link the observed outcome Y_i to the potential outcomes: $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. In addition, we assume that there are no unmeasured confounders, i.e. $\{Y(0), Y(1)\} \perp Z \mid \mathbf{X}$ and make the probabilistic assignment assumption: $0 < Pr(U = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x} \in \mathcal{P}$. The latter states that for any $\mathbf{x}$ in the target population $\mathcal{P}$, the probability of being assigned to concurrent RCT or historical control is bounded away from zero. Like SUTVA, these are also standard assumptions.

The causal estimand of the average treatment effect, conditional on $\mathbf{x}$, is defined as

$$\tau(\mathbf{x}) \equiv E(Y(1) - Y(0) \mid \mathbf{x}).$$

Let $f(\mathbf{x})$ denote the distribution of $\mathbf{x}$ in the target population $\mathcal{P}$, with respect to a product of counting measure for categorical variables and Lebegue measure on continuous variables. Our target causal estimand is the average treatment effect in $\mathcal{P}$, defined by averaging $\tau(\mathbf{x})$ over $f(\mathbf{x})$:

$$\tau_c \equiv \frac{\int \tau(\mathbf{x}) f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}. \tag{1}$$

The hypothesis of interest is $H_0 : \tau_c = 0$ vs $H_1 : \tau_c \neq 0$.

Let n_t, n_c, and n_h denote the sample sizes of the treatment arm, concurrent control arm, and historical control, respectively, and define $n = n_t + n_c + n_h$. We construct an estimator of τ_c as follows:

$$\hat{\tau}_c = \frac{\sum_i^n U_i Z_i Y_i}{\sum_i^n U_i Z_i} - \left\{ \pi \frac{\sum_i^n U_i (1 - Z_i) Y_i}{\sum_i^n U_i (1 - Z_i)} + (1 - \pi) \frac{\sum_i^n w_0(\mathbf{x}_i)(1 - U_i)(1 - Z_i) Y_i}{\sum_i^n w_0(\mathbf{x}_i)(1 - U_i)(1 - Z_i)} \right\}, \tag{2}$$

where the first term on the right-hand side is the estimator of the treatment effect of T. The second and third terms together (in the curly brackets) are the estimator of the treatment effect of C, taking a form of a weighted average of the estimator based on the concurrent control (i.e. the second term) and that based on the historical control (i.e. the third term) with weight $0 \leq \pi \leq 1$. We will discuss how to choose the weight π below. In the third term, the weights $w_0(\mathbf{x}_i)$ are used to map the historical control data from its source population $\mathcal{Q}$ to the trial target population $\mathcal{P}$. Making

causal inference across populations has been considered by Cole and Stuart (2010), Hartman et al. (2015), O'Muircheartaigh and Hedge (2014), Rudolph and van der Laan (2017), Dahabreh et al. (2019), and Li et al. (2018), among others.

We first discuss how to specify $w_0(\mathbf{x}_i)$. Define $s(\mathbf{x}_i) = \Pr(U = 1 \mid \mathbf{x}_i)$ as the propensity score (Rosenbaum & Rubin, 1983) for patient i, which is the probability of the patient belonging to the RCT given patient baseline covariates $\mathbf{x}_i$. $s(\mathbf{x}_i)$ is often estimated using a logistic regression based on data in both arms in the trial and historical control data. To reduce the impact of the misspecification of the propensity score model, a flexible logistic regression is preferred, e.g. a quadratic logistic regression with two-way interactions

$$\text{logit}(s(\mathbf{x})) = \eta_0 + \sum_j \alpha_j x_j + \sum_j \gamma_j x_j^2 + \sum_{i \neq j} \beta_{ij} x_i x_j. \tag{3}$$

Assuming that the propensity model is correctly specified, we have the following result with proof provided in the Appendix:

Theorem 1 $\hat{\tau}_c$ is a consistent estimator of τ_c when $w_0(\mathbf{x}_i) \propto \frac{s(\mathbf{x}_i)}{1-s(\mathbf{x}_i)}$.

The propensity score model requires correct specification of the covariates. Ideally, all important prognostic factors should be included, which requires careful discussion with subject-matter experts. One example is to include all baseline variables that the investigators plan on including in the primary analysis model. As the objective of propensity score weighting in this article is to balance covariate distributions in the sample, not to make inferences about assignment probabilities in the population, the propensity score model has a great tolerance to a large number of covariates and need not be parsimonious (Li et al., 2018). To reduce the impact of possibly extremely large values of $s(\mathbf{x}_i)$, trimming can be implemented to reduce variance of $\hat{\tau}_c$ (Elliott & Little, 2000; Lee et al., 2011; Potter, 1993).

We now discuss how to specify π. In equation (2), π controls the contribution of trial data and historical control data to estimate the treatment effect of C. A natural way is to weight the two estimates by their precisions, or equivalently their sample sizes. The challenge is that, as the estimate based on historical control data is obtained by mapping $\mathcal{Q}$ to $\mathcal{P}$, often only partial information is retained. Following McCaffrey et al. (2013), we measure the amount of information retained after the mapping using ESS:

$$\text{ESS}_H = \frac{\left(\sum_i^n w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)\right)^2}{\sum_i^n w_0^2(\mathbf{x}_i)(1-U_i)(1-Z_i)},$$

and accordingly set the weight π as

$$\pi = \frac{n_c}{n_c + \text{ESS}_H}.$$

As McCaffrey et al. (2013) pointed out, the weighted treatment effect from the historical control, $\sum_i^n w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)Y_i / \sum_i^n w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)$, generally has a greater sampling variance than unweighted means from a sample of equal size. ESS_H provides a useful measure for the potential loss in precision from weighting.

In the special case of the single-arm design without the concurrent control, the estimator of average treatment effect (2) reduces to

$$\hat{\tau}_c = \frac{\sum_i^n U_i Z_i Y_i}{\sum_i^n U_i Z_i} - \frac{\sum_i^n w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)Y_i}{\sum_i^n w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)}. \tag{4}$$

The standard error of $\hat{\tau}_c$ can be obtained using the non-parametric bootstrap (Efron & Tibshirani, 1994).

2.2 Trial design

The objective of the AHCC design is to emulate a standard RCT that randomises a total of N patients to T and C, referred to as the target RCT, where N is chosen based on the standard power calculation such that the target RCT has a reasonable (e.g. 80%) power to detect a prespecified effect size. For ease of exposition, we assume that the target RCT uses 1:1 randomisation, noting that the proposed method is applicable to targeting an RCT with any fixed ratio randomisation. Consider a K-stage AHCC design with $N_1, \ldots, N_K$ denoting the target cumulative sample sizes up to the end of stage k, $k = 1, \ldots, K$, with $N_K \equiv N$. For equal spaced K-stage design, $N_k = k(N/K)$. The AHCC design is described as follows:

1. At stage 1, enrol N_1 patients and randomise them in the ratio of $1:1$ to T and C.
2. At stage $k+1$, $k = 1, \ldots, K-1$, based on the cumulative interim data up to the end of stage k, we calculate ESS_H and $\hat{\tau}_c$ and perform futility and superiority interim analyses using O'Brien–Fleming boundary (O'Brien & Fleming, 1979). Specifically, we use a z test with the bootstrap standard error of $\hat{\tau}_c$. If the futility stopping boundary is crossed, terminate the trial and conclude that T is futile. If the superiority stopping boundary is crossed, terminate the trial and conclude that T is superior. Otherwise, calculate the total ESS (TESS, i.e. the number of concurrent patients + ESS_H) for the control arm C and randomise $(N_{k+1} - N_k)/2 + \max(n_0, N_{k+1}/2 - \text{TESS})$ patients in the ratio of $[(N_{k+1} - N_k)/2] : [\max(n_0, N_{k+1}/2 - \text{TESS})]$ to T and C, where n_0 is a prespecified small integer, representing the minimal number of patients to be randomised to C in a stage. The randomisation ratio is chosen to emulate the target 1:1 RCT, for which the sample size of T and C is $N_{k+1}/2$ after stage $k+1$. Note that N_k is the target/planned cumulative sample size at the end of stage k, not the actual randomised sample size, so does not depend on the observed data. We use n_0 to ensure that even when the historical control provides more ESS than what stage $k+1$ needs (i.e. $\text{TESS} \geq N_{k+1}/2$), patient assignment still involves a certain degree of randomisation. n_0 can be set as approximately the number of patients to be randomised to C to ensure a small randomisation probability, say 5%, to C. An illustrated example is provided in the Supplementary Materials.
3. At the end of the trial, we test the hypothesis $H_0{:}\tau_c = 0$ vs $H_1 : \tau_c \neq 0$ based on all the trial data and historical control data. If the p value is less than the O'Brien–Fleming boundary, H_0 is rejected; otherwise, we fail to reject H_0.

In Step 2, updating the randomisation ratio and the interim futility/superiority analyses are not binding and can be planned at different frequencies. That is, at a certain stage, we can just update the randomisation ratio without performing futility/superiority analyses, or vice versa. In practice, due to logistic considerations, the two-stage design may be the most likely choice. Thus, in the simulations, we mainly focus on the two-stage AHCC design. In addition, we employ the O'Brien–Fleming boundary for superiority and futility stopping. Other alpha/beta spending functions (e.g. DeMets & Lan, 1994) can also be used to provide additional flexibility, e.g. no need to pre-specify the interim time and frequency.

2.2.1 Design with single-arm stage I

A variation or limiting case of the above design is to start the trial with a single-arm design by assigning all the initial stage's patients to the treatment arm. Based on the interim data, randomisation may be initiated in later stages when needed (e.g. $\text{ESS}_H < N_k/2$ in Step 2). Simulation shows that this approach generally has similar operating characteristics as the above design (see Supplementary Materials), but it provides extra sample size saving if $\text{ESS}_H > N_k/2$. Therefore, it may be advantageous if there is strong prior evidence that the historical control population is comparable to the trial population and thus substantial information can be borrowed (e.g. if the historical data are very recent and the patient eligibility criteria are expected to be the same for historical data and current RCT), such that randomisation of patients to the concurrent control is not needed. Nevertheless, in many cases, regulatory agencies prefer some concurrent control data to facilitate the evaluation of the treatment effect. Therefore, we focus on the AHCC design starting with randomisation.

3 Simulation study

3.1 Continuous outcome

We conducted simulation studies to evaluate the operating characteristics of the AHCC design. We assumed the covariates $\mathbf{x}$ with dimension $p = 3$ for patients in the current RCT followed a multivariate normal distribution $\phi_p(\mu_1, \Sigma_1)$, where $\mu_1 = (1, 1, 1)$, and

$$\Sigma_1 = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}.$$

We generated Y_i from model

$$Y_i \mid \mathbf{x}_i, Z_i = \beta_0 + \tau Z_i + \mathbf{x}_i\beta + \epsilon_i, \tag{5}$$

where $\beta_0 = 0, \beta = (1, 1, 1), \epsilon_i \sim N(0, 1)$, and $\tau = 0$ (for evaluating type I error) or 0.905 (for evaluating power). Under this setting, the standard deviation (sd) under each treatment arm is 2.28. For a standard RCT using a two-tailed two-sample t-test, 100 patients are needed in each treatment arm to achieve 80% power at true treatment effect $\tau = 0.905$ when the sd is 2.28 for each arm, so we set $N = 200$. This design is what we referred to earlier as the targeted RCT.

We set N_0, the number of historical control patients, to be 500. For historical control patients, we also assumed the covariates $\mathbf{x}$ followed a multivariate normal distribution $\phi_p(\mu_0, \Sigma_0)$, where μ_0 and Σ_0 were chosen to achieve varying ESS for historical data. Details are provided in the Supplementary Materials. We constructed 14 scenarios with ESS ranging from 3 to 140, as listed in Table 1. In particular, the scenario with ESS = 3 was used to examine the situation where the historical data contributed almost no information. For historical control data, we generated Y_i from the same model as in RCT

$$Y_i \mid \mathbf{x}_i = \beta_0 + \mathbf{x}_i\beta + \epsilon_i. \tag{6}$$

We used 200 bootstrap samples to estimate the standard error of $\hat{\tau}_c$, which is generally a sufficient sample number for estimating standard errors (Efron & Tibshirani, 1994). The propensity score $s(\mathbf{x})$ was estimated using a quadratic logistic regression with two-way interactions in equation (3) for all simulation studies including the sensitivity analyses. We implemented the two-stage AHCC design with the first-stage sample size $N_1 = 120$, which is 60% information fraction. We compared it to the standard 1:1 RCT that has the maximum sample size of 200 patients with an interim after 120 patients are randomised. The O'Brien–Fleming futility and superiority stopping rules were applied to the two designs at the interim, and the EAST software was used to obtain the stopping boundaries. The interim α value is 0.008 for superiority and 0.466 for futility; and at the end of the trial, the α value is 0.05. We evaluated the type I error at $\tau = 0$ and power at $\tau = 0.905$. The minimal number of patients to be randomised to C in stage II was taken as $n_0 = 4$ so that approximately 5% of patients would be assigned to C. Trimming was implemented by trimming patients in both arms with propensity score values less than 0.1 or greater than 0.9. For each scenario, we simulated 10,000 trials.

Table 1 shows the operating characteristics of the two-stage AHCC design with interim sample size 120 across 10,000 simulated trials for the 14 scenarios with ESS ranging from 3 to 140. Table 1 reports the type I error, power at $\tau = 0.905$, numbers of patients assigned to the RCT treatment and concurrent control arms, percentage of average sample size saved compared with a standard two-stage RCT, and the percentages of early stopping due to superiority and futility. We also report the relative bias of $\hat{\tau}_c$, which is defined as the difference between the absolute empirical bias of the standard RCT and that of $\hat{\tau}_c$. As shown by Table 1, the AHCC design controlled the type I error rate around 5% across all ESS, and the power at $\tau = 0.905$ was greater than 80%. The relative bias is mostly positive for the 14 scenarios, indicating that the bias of the AHCC design was slightly smaller than that of the standard RCT. As ESS increased from 3 to 140, power increased from 0.825 to 0.973; the sample size of the concurrent control arm decreased from 83.3 (when $\tau = 0$) and 80.3 (when $\tau = 0.905$) to 61; the percentage of sample size saving increased from

Table 1. Continuous outcome: type I error, power, relative bias of $\hat{\tau}_c$, percentages of early stopping due to superiority and futility for the two-stage AHCC design with interim sample size 120

$\tau = 0$								
ESS	**Type I error**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.054	89.2	83.3	172.5	3.3	0.002	0.4	26.5
10	0.053	89.7	76.9	166.6	6.7	−0.011	0.5	25.1
20	0.054	88.9	70.2	159.1	10.9	0.005	0.5	27.1
30	0.052	88.9	65.4	154.3	13.5	0.008	0.5	26.7
40	0.052	88.1	62.4	150.4	15.7	0.001	0.4	27.4
50	0.052	87.9	61.6	149.5	16.2	0.010	0.6	26.6
60	0.047	87.6	61.5	149.0	16.5	0.007	0.5	27.0
70	0.055	87.7	61.5	149.1	16.5	0.005	0.6	26.6
80	0.052	87.7	61.5	149.2	16.4	0.004	0.5	26.6
90	0.049	87.8	61.5	149.3	16.4	0.006	0.5	26.3
100	0.048	87.6	61.4	149.0	16.5	0.007	0.3	27.1
110	0.047	87.5	61.5	149.0	16.5	0.002	0.3	27.2
125	0.053	87.7	61.5	149.1	16.5	0.001	0.5	26.6
140	0.048	87.8	61.5	149.3	16.4	0.005	0.4	26.3
$\tau = 0.905$								
ESS	**Power**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.825	85.4	80.3	165.7	4.8	−0.011	35.7	0.9
10	0.858	82.9	73.0	155.9	10.4	−0.021	42.1	0.7
20	0.856	82.1	67.9	150.1	13.8	0.002	44.2	0.5
30	0.863	80.9	64.1	144.9	16.8	0.013	47.0	0.5
40	0.871	80.0	61.7	141.7	18.6	0.032	48.2	0.5
50	0.891	78.4	61.1	139.5	19.9	0.019	51.7	0.3
60	0.905	77.4	60.9	138.3	20.6	0.026	54.0	0.4
70	0.916	76.8	60.9	137.7	20.9	0.035	55.5	0.2
80	0.928	75.3	60.8	136.1	21.8	0.030	59.4	0.2
90	0.942	74.8	60.8	135.6	22.1	0.028	61.0	0.1
100	0.950	74.3	60.8	135.1	22.4	0.031	62.1	0.1
110	0.959	73.5	60.7	134.2	22.9	0.033	64.4	0.1
125	0.966	72.6	60.7	133.3	23.4	0.036	66.6	0.2
140	0.973	72.1	60.6	132.8	23.7	0.034	68.0	0.0

Note. n_t is the number of patients in the treatment arm; n_c is the number of patients in the concurrent control arm; n is the total number of patients in the trial; % n save is the percentage of sample size saved for the trial by using AHCC design leveraging historical data compared with a standard two-stage RCT.

3.3% to 16.4% when $\tau = 0$ and from 4.8% to 23.7% when $\tau = 0.905$. It is of interest that the power of AHCC increased with ESS even before it reached 100 and was often higher than 80%. This power gain is due to the fact that AHCC incorporates the historical data by mapping it to the current trial data through propensity score weighting. The mapped historical data are more similar to the trial data, compared to the complete randomness in the randomised control, thereby increasing power. These results suggest that compared with the standard RCT, leveraging historical data may dramatically reduce the sample size and increase power when the historical data ESS is large. When there is little information to borrow from historical data (e.g. when ESS = 3), the performance of the AHCC design is similar to that of the standard RCT.

3.2 Binary outcome

We also considered binary outcome Y. The same 14 scenarios of covariate distributions as in the continuous outcome case with ESS ranging from 3 to 140 were used. We simulated binary Y_i from the logistic model

$$\text{logit}(P(Y_i = 1 \mid \mathbf{x}_i, Z_i)) = \beta_0 + \tau Z_i + \mathbf{x}_i\beta.$$

We set $\beta = (1, 1, 1)$. For a two-sided two-sample proportion test, a sample size of 100 in each treatment arm would achieve an 80% power when the two proportions are 0.2 and 0.378. So, we chose $\beta_0 = -5.25$ and $\boldsymbol{\tau} = \mathbf{1.417}$ so that $Pr(Y = 1 \mid Z = 1) = 0.378$ and $Pr(Y = 1 \mid Z = 0) = 0.2$. We evaluated type I error at $\tau = 0$ and power at $\tau = 1.417$.

Table 2 summarises the operating characteristics of the two-stage AHCC design for binary outcome with an interim sample size of 120. As Table 2 shows, the results are very similar to those for the continuous outcomes shown in Table 1. When there is little information to borrow from historical control data (i.e. when ESS = 3), the AHCC design yielded similar results as the standard RCT. When historical control data provide more information (because they are more compatible to the current trial data), the AHCC design resulted in reduced sample size, with similar or higher power.

3.3 Sensitivity analysis

Operating characteristics of the three-stage AHCC design with target interim sample sizes 70 and 135 are displayed in Tables 3 and 4 for continuous and binary outcomes, respectively. Given the possibly immature data involved for efficacy conclusion, only futility stopping was implemented at the first interim analysis, while both efficacy and futility stopping were implemented at the second interim analysis. Operating characteristics of the two-stage AHCC design with interim sample sizes 60 and 100 are provided in online supplementary material, Table S1–S4. We also assessed the performance of the AHCC design when the randomisation ratio is 3:1 with one interim analysis at 120 patients. Operating characteristics are shown in online supplementary material, Table S7. Under all these settings, the AHCC design improves power and reduces sample size when ESS is greater than 10, compared with standard RCT.

We also evaluated the performance of the AHCC design when starting the trial with a single-arm design by assigning all stage 1 patients to the treatment arm. Results are displayed in online supplementary material, Table S5 (continuous outcome) and online supplementary material, Table S6 (binary outcome). The AHCC design provided similar operating characteristics as in the aforementioned settings. However, the percentage of sample size saved reached over 31% when ESS was greater than 100, as compared to the counterpart of 16.4–23.7% (Tables 1 and 2) when starting the trial with randomisation.

We performed additional simulation studies to evaluate the robustness of the AHCC design in the situation of missed confounders. We assumed four covariates affected the outcome Y, so $p = 4$, while only three covariates were included for analysis. We assumed the covariates $\mathbf{x}$ with dimension $p = 4$ for patients in the current RCT followed a multivariate normal distribution $\phi_p(\mu_1, \Sigma_1)$, where $\mu_1 = (1, 1, 1, 1)$, and

$$\Sigma_1 = \begin{pmatrix} 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 \end{pmatrix}.$$

We generated Y_i from model

$$Y_i \mid \mathbf{x}_i, Z_i = \beta_0 + \tau Z_i + \mathbf{x}_i\beta + \epsilon_i, \tag{7}$$

where $\beta_0 = 0$, $\beta = (1, 1, 1, \beta_4)$, $\epsilon_i \sim N(0, 1)$. Three values of β_4, (0.5, 1, 2), were used to evaluate the effect of the missed confounder, x_4, on the performance of AHCC. The standard deviation

Table 2. Binary outcome: type I error, power, relative bias, percentages of early stopping due to superiority and futility for the two-stage AHCC design with interim sample size 120

$\tau = 0$								
ESS	**Type I error**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.056	89.3	83.3	172.6	4.6	0.000	0.4	26.5
10	0.054	89.4	76.7	166.1	8.2	0.002	0.5	26.1
20	0.048	89.4	70.4	159.9	11.7	0.002	0.4	26.1
30	0.051	89.2	65.5	154.7	14.5	0.002	0.4	26.2
40	0.053	88.7	62.4	151.1	16.5	0.001	0.3	25.9
50	0.050	87.8	61.7	149.5	17.4	0.002	0.3	27.1
60	0.052	87.8	61.5	149.3	17.5	0.002	0.4	26.5
70	0.053	87.9	61.5	149.4	17.5	0.002	0.2	26.4
80	0.049	87.9	61.5	149.4	17.5	0.002	0.3	26.4
90	0.054	88.0	61.5	149.5	17.4	0.002	0.2	26.2
100	0.053	87.8	61.5	149.3	17.5	0.002	0.3	26.6
110	0.050	87.9	61.5	149.3	17.5	0.002	0.2	26.5
125	0.049	87.7	61.4	149.2	17.6	0.002	0.2	26.8
140	0.049	87.8	61.5	149.3	17.5	0.001	0.2	26.7
$\tau = 1.417$								
ESS	**Power**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.802	86.2	80.9	167.0	4.0	0.003	33.5	1.1
10	0.808	84.7	74.0	158.7	8.8	0.002	37.2	1.1
20	0.816	83.9	68.5	152.4	12.4	0.003	39.3	0.9
30	0.819	83.3	64.4	147.7	15.1	0.002	40.6	0.6
40	0.827	82.1	61.9	144.0	17.3	0.004	42.4	0.8
50	0.833	82.0	61.3	143.4	17.6	0.003	41.9	0.6
60	0.846	81.4	61.1	142.6	18.1	0.002	43.2	0.5
70	0.869	80.3	61.1	141.3	18.8	0.002	46.4	0.4
80	0.887	79.5	61.0	140.5	19.3	0.002	48.3	0.4
90	0.885	79.7	61.0	140.7	19.1	0.003	47.9	0.3
100	0.899	79.0	61.0	140.0	19.5	0.003	49.5	0.4
110	0.899	78.9	61.0	139.8	19.6	0.002	50.0	0.4
125	0.913	78.4	61.0	139.4	19.9	0.003	51.2	0.3
140	0.915	78.4	61.0	139.3	19.9	0.003	51.5	0.3

Note. n_t is the number of patients in the treatment arm; n_c is the number of patients in the concurrent control arm; n is the total number of patients in the trial; % n save is the percentage of sample size saved for the trial by using AHCC design leveraging historical data compared with a standard two-stage RCT.

of each arm under current trial is 2.46, 2.72, 3.4, respectively, when $\beta_4 = 0.5, 1, 2$. For a standard two-tailed two-sample t-test for an RCT, a sample size of 100 in each arm is required to achieve an 80% power at $\tau = 0.978, 1.08, 1.35$, for $\beta_4 = 0.5, 1, 2$, respectively.

For historical control patients, we constructed three scenarios assuming the covariates $\mathbf{x}$ follow a multivariate normal distribution $\phi_p(\mu_0, \Sigma_0)$, where μ_0 and Σ_0 (provided in the Supplementary Materials) were chosen so that the ESS was 30 for historical data with sample size $N_0 = 500$

Table 3. Continuous outcome: type I error, power, relative bias, percentages of early stopping due to superiority and futility of $\hat{\tau}_c$ for the three-stage AHCC design with target interim sample sizes 70 and 135

$\tau = 0$								
ESS	**Type I error**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.054	87.4	80.0	167.4	3.8	0.008	0.7	34.7
10	0.054	88.1	72.5	160.7	7.7	−0.007	0.7	32.8
20	0.053	88.0	64.5	152.5	12.4	0.001	0.8	33.4
30	0.053	87.6	57.3	144.9	16.7	0.006	0.8	33.8
40	0.051	87.0	50.5	137.5	21.0	−0.008	0.6	35.0
50	0.051	87.0	45.4	132.4	23.9	0.009	0.7	34.4
60	0.051	86.4	41.0	127.4	26.8	0.009	0.7	35.1
70	0.047	86.0	38.8	124.7	28.3	0.004	0.7	35.3
80	0.048	85.8	37.9	123.7	28.9	0.000	0.7	34.9
90	0.054	86.2	37.7	124.0	28.8	0.008	0.7	33.8
100	0.047	85.6	37.7	123.3	29.2	−0.002	0.5	35.7
110	0.049	85.7	37.7	123.3	29.1	0.001	0.6	35.3
125	0.050	85.5	37.7	123.2	29.2	0.001	0.6	35.9
140	0.048	85.8	37.7	123.4	29.1	0.008	0.5	35.1
$\tau = 0.905$								
ESS	**Power**	n_t	n_c	n	% n save	**Relative bias**	**% super**	**% futil**
3	0.827	84.8	77.6	162.4	4.8	−0.036	44.5	1.1
10	0.856	83.2	68.0	151.1	11.4	−0.042	49.9	0.8
20	0.860	82.7	59.4	142.1	16.7	−0.019	51.0	0.8
30	0.869	81.7	52.2	133.9	21.5	−0.011	53.5	0.7
40	0.879	80.1	46.1	126.2	26.0	0.006	57.4	0.6
50	0.898	79.1	42.4	121.4	28.8	−0.002	60.0	0.4
60	0.920	76.9	39.2	116.2	31.9	0.000	65.6	0.4
70	0.931	75.8	37.8	113.6	33.4	0.018	68.5	0.3
80	0.945	75.1	37.2	112.3	34.2	0.023	70.4	0.3
90	0.957	73.9	37.1	111.0	34.9	0.020	74.0	0.1
100	0.968	73.0	37.0	110.0	35.5	0.020	76.5	0.2
110	0.969	72.3	37.0	109.3	35.9	0.018	78.6	0.1
125	0.979	71.7	36.9	108.7	36.3	0.024	80.6	0.1
140	0.983	71.0	36.9	107.8	36.8	0.023	83.0	0.1

Note. n_t is the number of patients in the treatment arm; n_c is the number of patients in the concurrent control arm; n is the total number of patients in the trial; % n save is the percentage of sample size saved for the trial by using AHCC design leveraging historical data compared with a standard three-stage RCT.

when mapped to a current RCT dataset of size 100. Historical control data were generated from the same model as in RCT, i.e.

$$Y_i \mid \mathbf{x}_i = \beta_0 + \mathbf{x}_i\beta + \epsilon_i. \tag{8}$$

We implemented the AHCC design with the first three x variables ignoring x_4. Table 5 summarises the performance of the AHCC design. Across the three values of β_4, the type I error was controlled

Table 4. Binary outcome: type I error, power, relative bias, percentages of early stopping due to superiority and futility of $\hat{\tau}_c$ for the three-stage AHCC design with target interim sample sizes 70 and 135

$\tau = 0$								
ESS	**Type I error**	n_t	n_c	n	**% n save**	**Relative bias**	**% super**	**% futil**
3	0.050	87.7	80.4	168.1	6.4	−0.001	0.7	34.3
10	0.053	87.5	71.9	159.4	11.2	0.000	0.8	34.6
20	0.056	87.8	64.3	152.1	15.3	−0.002	0.9	33.5
30	0.052	87.2	57.2	144.5	19.6	−0.002	0.8	34.7
40	0.058	87.3	50.5	137.8	23.2	−0.001	0.8	34.2
50	0.054	86.7	45.3	132.0	26.5	0.001	0.8	34.9
60	0.053	86.9	41.0	127.8	28.8	0.001	0.8	33.6
70	0.047	86.3	38.8	125.1	30.3	0.000	0.5	34.5
80	0.048	86.0	37.9	123.9	31.0	0.000	0.5	34.8
90	0.047	86.2	37.7	124.0	31.0	0.000	0.4	34.0
100	0.049	86.5	37.7	124.2	30.8	0.001	0.5	33.3
110	0.045	85.9	37.7	123.6	31.2	0.001	0.5	34.7
125	0.052	86.1	37.7	123.8	31.1	0.000	0.5	34.1
140	0.047	85.9	37.7	123.6	31.2	0.000	0.4	34.9
$\tau = 1.417$								
ESS	**Power**	n_t	n_c	n	**% n save**	**Relative bias**	**% super**	**% futil**
3	0.800	85.3	78.1	163.4	5.2	0.001	42.3	1.7
10	0.812	84.7	69.3	153.9	10.7	0.000	44.9	1.2
20	0.811	84.6	61.2	145.8	15.4	0.001	44.6	1.3
30	0.829	83.8	53.9	137.7	20.1	−0.001	46.7	1.2
40	0.830	82.8	47.7	130.5	24.3	0.001	49.1	0.8
50	0.819	82.6	43.6	126.2	26.8	0.000	48.9	1.1
60	0.823	82.0	40.2	122.2	29.1	0.002	50.0	0.9
70	0.863	79.9	38.1	118.0	31.5	0.002	55.8	0.6
80	0.884	79.5	37.5	117.0	32.1	0.002	56.9	0.4
90	0.881	78.9	37.4	116.3	32.5	0.003	58.2	0.6
100	0.893	78.7	37.3	116.0	32.7	0.003	58.9	0.5
110	0.903	78.1	37.3	115.4	33.1	0.001	60.9	0.4
125	0.914	77.7	37.2	114.9	33.3	0.002	62.2	0.4
140	0.927	77.0	37.2	114.3	33.7	0.002	64.0	0.4

Note. n_t is the number of patients in the treatment arm; n_c is the number of patients in the concurrent control arm; n is the total number of patients in the trial; % n save is the percentage of sample size saved for the trial by using AHCC design leveraging historical data compared with a standard three-stage RCT.

at 5% for scenario 1, while between 0.055 and 0.088 for scenario 2, and between 0.058 and 0.101 for scenario 3. As β_4 increased from 0.5 to 2, power decreased from 0.881 to 0.791 for scenario 1, but it was about the same for scenarios 2 and 3. The relative bias increased with β_4 for scenarios 2 and 3.

The missed confounder x_4 had minor to moderate impact on the performance of the AHCC design for these three scenarios. Missed confounders can have substantial impact on type I error and/or power, therefore it is important to carefully identify and include all potential confounders in the propensity model.

Table 5. Sensitivity analysis: Type I error and power for three scenarios with 'true' ESS = 30

	$\tau = 0$					$\tau = \tau_0$				
β_4	Type I error	n_t	n_c	n	Relative bias	Power	n_t	n_c	n	Relative bias
Scenario 1										
0.5	0.052	87.9	61.7	149.6	0.011	0.881	79.2	61.2	140.4	0.019
1	0.049	87.6	61.6	149.2	0.005	0.852	80.4	61.2	141.6	0.031
2	0.053	87.5	61.6	149.2	0.009	0.791	82.6	61.4	144.0	0.060
Scenario 2										
0.5	0.055	88.7	61.7	150.4	−0.037	0.911	76.5	61.0	137.6	−0.049
1	0.069	89.4	61.8	151.2	−0.102	0.921	75.8	61.0	136.7	−0.103
2	0.088	89.9	61.8	151.7	−0.204	0.907	75.6	61.0	136.6	−0.208
Scenario 3										
0.5	0.058	88.6	61.6	150.2	−0.047	0.918	75.8	60.9	136.7	−0.049
1	0.071	89.3	61.6	150.9	−0.114	0.931	74.7	60.9	135.6	−0.125
2	0.101	90.0	61.7	151.7	−0.253	0.925	74.2	60.8	135.1	−0.264

Note. n_t is the number of patients in the treatment arm; n_c is the number of patients in the concurrent control arm; n is the total number of patients in the trial. Type I error is evaluated at $\tau = 0$, and power is evaluated at $\tau_0 = 0.978$, 1.08, 1.35 for $\beta_4 = 0.5$, 1, 2, respectively.

4 Discussion

We have proposed the AHCC design to leverage historical control data into an RCT to reduce the sample size. Under the estimand framework, we explicitly define the causal estimand of the average treatment effect, construct an estimator of the estimand, and prove that the estimator is a consistent estimator of the estimand. The effective sample size is used to quantify the contribution of the historical data. The AHCC design takes a multistage approach. Simulation studies show our method generally has improved power and saves on sample size compared with standard RCTs.

In this article, we consider a two-arm RCT. The approach is readily extendable to multiple-arm RCTs in a straightforward way. In our simulation studies, we used the logistic regression to estimate the propensity score, which demonstrated its sufficiency for a wide range of scenarios as the operating characteristics showed. As commented by Li et al. (2018), in practice, it is important to calibrate the propensity score model, and a richer model, e.g. the generalised boosted model (McCaffrey et al., 2004), may be desirable in certain situations.

One limitation of our method is that it is applicable mainly to the case of population drift of patient's baseline covariates. It cannot efficiently account for other time trends, such as changes in the outcomes over time possibly due to changes in standard care and/or learning curves amongst study personnel, changes in the disease itself (e.g. new variants), and seasonal effects. In these cases, one approach is to implement a weighted regression with a treatment effect and a smooth time effect using, e.g. spline models (Eilers & Marx, 1996; Green & Silverman, 1993; Ruppert et al., 2003). The weight for historical control patient i depends on the weight $w_0(\mathbf{x}_i)$ to balance the covariate distributions of the concurrent RCT and historical control patients. Also in this article, we consider continuous or binary outcome. For time-to-event outcomes, we can extend the methods proposed in Austin (2014), Austin and Stuart (2017), and Zhang and Kim (2019), which use matching, weighting, stratification, or regression. These are areas of our future research.

Since our method depends on the correct specification of the propensity score model, calibration of the propensity score model can be performed by comparing the predicted

and observed rates of the assignment. If miscalibration is identified, more flexible models such as spline models or ad hoc methods such as ratio adjustment of each decile can be used (Li et al., 2018).

Acknowledgments

The authors thank the Associate Editor and three Referees for their valuable comments which substantially improved the presentation of this article. B.G.'s research is supported by the R & D Research Competitiveness Subprogram of Louisiana Board of Regents, Contract number LEQSF(2022-25)-RD-A-07. Y.Y.'s research was partially supported by Award Number P50CA281701, P50CA221707, and P50CA127001 from the National Cancer Institute.

Conflict of interest: None declared.

Data availability

No new data were generated or analysed as part of this research.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

Appendix

Proof of Theorem 1. Under the unconfoundedness assumption, the causal estimand is defined as

$$\begin{aligned}\tau_c &= \frac{\int \tau(\mathbf{x}) f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} \\ &= \frac{\int E_{Y,Z,U|\mathbf{x}} \left\{ Y(1)ZU - Y(0)(1-Z)U \right\} f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} \\ &= \frac{\int E_{Y,Z,U|\mathbf{x}} Y(1)ZU f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} ZU f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} - \frac{\int E_{Y,Z,U|\mathbf{x}} Y(0)(1-Z)U f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} (1-Z)U f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\tau_c &= \frac{\int \tau(\mathbf{x}) f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} \\ &= \frac{\int E_{Y,Z,U|\mathbf{x}} \left\{ Y(1)ZU - Y(0)(1-Z)(1-U)[s(\mathbf{x})/(1-s(\mathbf{x}))] \right\} f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} \\ &= \frac{\int E_{Y,Z,U|\mathbf{x}} Y(1)ZU f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} ZU f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})} \\ &\quad - \frac{\int E_{Y,Z,U|\mathbf{x}} Y(0)(1-Z)(1-U)[s(\mathbf{x})/(1-s(\mathbf{x}))] f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} (1-Z)(1-U)[s(\mathbf{x})/(1-s(\mathbf{x}))] f(\mathbf{x})/s(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})}.\end{aligned}$$

Letting

$$\tau_1 = \frac{\int E_{Y,Z,U|\mathbf{x}} Y(1) Z U f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} Z U f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})},$$

$$\tau_{01} = \frac{\int E_{Y,Z,U|\mathbf{x}} Y(0)(1-Z) U f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} (1-Z) U f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})},$$

$$\tau_{00} = \frac{\int E_{Y,Z,U|\mathbf{x}} Y(0)(1-Z)(1-U)[s(\mathbf{x})/(1-s(\mathbf{x}))] f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})}{\int E_{Z,U|\mathbf{x}} (1-Z)(1-U)[s(\mathbf{x})/(1-s(\mathbf{x}))] f(\mathbf{x})/s(\mathbf{x})\mu(\mathrm{d}\mathbf{x})},$$

we have $\tau_{00} = \tau_{01}$. Let $\tau_0 = \tau_{00} = \tau_{01}$, then $\tau_c = \tau_1 - \tau_0$. τ_1 is the expected value of the means of Y in samples drawn from the treatment arm in the current RCT; τ_{01} is the expected value of the mean of Y in samples drawn from the concurrent arm in the current RCT; and τ_{00} is the expected value of the weighted means of Y in samples drawn from historical control data. Replacing expectations by sample means and with $w_0(\mathbf{x}_i) = \frac{s(\mathbf{x}_i)}{1-s(\mathbf{x}_i)}$ defined in (3), $\frac{\sum_i U_i Z_i Y_i}{\sum_i U_i Z_i}$ is an unbiased and consistent estimator of τ_1; $\frac{\sum_i U_i (1-Z_i) Y_i}{\sum_i U_i (1-Z_i)}$ is an unbiased and consistent estimator of τ_{01}, and $\frac{\sum_i w_0(\mathbf{x}_i)(1-U_i)(1-Z_i) Y_i}{\sum_i w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)}$ is an unbiased and consistent estimators of τ_{00}. So when p is between 0 and 1, $p \frac{\sum_i U_i (1-Z_i) Y_i}{\sum_i U_i (1-Z_i)} + (1-p) \frac{\sum_i w_0(\mathbf{x}_i)(1-U_i)(1-Z_i) Y_i}{\sum_i w_0(\mathbf{x}_i)(1-U_i)(1-Z_i)}$ is an unbiased and consistent estimator of τ_0. By Slutsky's theorem, $\hat{\tau}_c$ is a consistent estimator of τ_c. □

References

Austin P. (2014). The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, *33*(7), 1242–1258. https://doi.org/10.1002/sim.v33.7

Austin P., & Stuart E. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*, *26*(4), 1654–1670. https://doi.org/10.1177/0962280215584401

Berry S. M., Broglio K. R., Groshen S., & Berry D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials*, *10*(5), 720–734. https://doi.org/10.1177/1740774513497539

Chen W. C., Wang C., Li H., Lu N., Tiwari R., Xu Y., & Yue L. Q. (2020). Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*, *30*(3), 508–520. https://doi.org/10.1080/10543406.2020.1730877

Chu Y., & Yuan Y. (2018a). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials*, *15*(2), 149–158. https://doi.org/10.1177/1740774518755122

Chu Y., & Yuan Y. (2018b). BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *Journal of the Royal Statistical Society: Series C*, *67*(3), 723–740. https://doi.org/10.1111/rssc.12255

Cole S. R., & Stuart E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Dahabreh I., Robertson S., Tchetgen E., Stuart E., & Herman M. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, *75*(2), 685–694. https://doi.org/10.1111/biom.v75.2

DeMets D. L., & Lan K. K. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, *13*(13-14), 1341–1352. https://doi.org/10.1002/sim.v13:13/14

Efron B., & Tibshirani R. (1994). *An introduction to the bootstrap* (p. 52). Chapman & Hall.

Eilers P., & Marx B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102. https://doi.org/10.1214/ss/1038425655

Elliott M., & Little R. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, *16*, 191–209.

Green P., & Silverman B. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman and Hall.

Hartman E., Grieve R., Ramsahai R., & Sekhon J. S. (2015). From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A*, *178*, 757–779. https://doi.org/10.1111/rssa.12094

Hobbs B. P., Carlin B. P., Manderkar S. J., & Sargent D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, *67*(3), 1047–1056. https://doi.org/10.1111/biom.2011.67.issue-3

Ibrahim J. G., & Chen M. H. (2000). Power prior distribution for regression models. *Statistical Science*, *15*, 46–60. https://doi.org/10.1214/ss/1009212673

Ibrahim J. G., Chen M. H., & Sinha D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association*, *98*(461), 204–13. https://doi.org/10.1198/016214503388619229

Imbens G. W., & Rubin D. B. (2015). *Causal inference for statistics, social and biomedical science: An introduction*. Cambridge University Press.

Jiang L., Nie L., & Yuan Y. (2023). Elastic priors to dynamically borrow information from historical data in clinical trials. *Biometrics*, *79*(1), 49–60. https://doi.org/10.1111/biom.v79.1

Lee B., Lessler J., & Stuart E. (2011). Weight trimming and propensity score weighting. *PLoS One*, *6*(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Li F., Morgan K., & Zaslavsky A. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li H. F., Chen W. C., Wang C., Lu N., Song C., Tiwari R., Xu Y., & Yue L. Q. (2021). Augmenting both arms of a randomized controlled trial using external data: An application of the propensity score-integrated approaches. *Statistics in Biosciences*, *19*, 1–11. https://doi.org/10.1007/s12561-021-09315-5

Lin J., Gamalo-Seibers M., & Tiwari R. (2019). Propensity-score-based priors for Bayesian augmented control design. *Pharmaceutical Statistics*, *18*(2), 223–238. https://doi.org/10.1002/pst.v18.2

Lu N., Wang C., Chen W., Li H., Song C., Tiwaqri R., Xu Y., & Yue L. (2022). Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources. *Journal of Biopharmaceutical Statistics*, *32*(1), 158–169. https://doi.org/10.1080/10543406.2021.1998098

McCaffrey D., Griffin B., Almirall D., Slaughter M., Ramchand R., & Burgette L. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388–3414. https://doi.org/10.1002/sim.v32.19

McCaffrey D., Ridgeway G., & Morral A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

O'Brien P. C., & Fleming T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–56. https://doi.org/10.2307/2530245

O'Muircheartaigh C., & Hedge L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society Series C*, *63*(2), 195–210. https://doi.org/10.1111/rssc.12037

Pfizer (2019). US Food and Drug Administration approves Ibrance (Palbociclib) for the treatment of men with HR+, HER2-Metastatic breast cancer [www.pfizer.Com]. https://www.pfizer.com/news/press-release/press-release-detail/u˙s˙fda˙approves˙ibrance˙palbociclib˙for˙the˙treatment˙of˙men˙with˙hr˙her2˙metastatic˙breast˙cancer!.

Potter F. J. (1993). The effect of weight trimming on nonlinear survey estimates. *Journal of the American Statistical Association*, *2*, 758–763.

Przepiorka D., Ko C. W., Deisseroth A., Yancey C. L., Candau-Chacon R., Chiu H. J., Gehrke B. J., Gomez-Broughton C., Kane R. C., Kirshner S., & Mehrotra N. (2015). FDA approval: Blinatumomab. *Clinical Cancer Research*, *21*(18), 4035–4039. https://doi.org/10.1158/1078-0432.CCR-15-0612

Rosenbaum P. R., & Rubin D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *The Journal of the Royal Statistical Society, B*, *45*, 212–218. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58. https://doi.org/10.1214/aos/1176344064

Rubin D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, *75*, 591–593. https://doi.org/10.2307/2287653

Rudolph K. E., & van der Laan M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *The Journal of the Royal Statistical Society, B*, *79*(5), 1509–1525. https://doi.org/10.1111/rssb.12213

Ruppert D., Wand M., & Carroll R. (2003). *Semiparametric regression*. Cambridge University Press.

Stuart E., & Rubin D. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279–306. https://doi.org/10.3102/1076998607306078

Thall P. F., Wathen J. K., Bekele B. N., Champlin R. E., Baker L. H., & Benjamin R. S. (2003). Hierarchical Bayesian approaches to phase II trials in disease with multiple subtypes. *Statistics in Medicine*, *22*(5), 763–780. https://doi.org/10.1002/(ISSN)1097-0258

US Food and Drug Administration (2017a). *FDA approves first treatment for a form of batten disease [press release]*. Online.

US Food and Drug Administration (2017b). *Use of real-world evidence to support regulatory decision-making for medical devices*. Online.

US Food and Drug Administration (2021). *Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/ submitting-documents-using-real-worlddata-and-real-world-evidence-fda-drugsand-biologics-guidance!.

Wang C., Li H., Chen W. C., Lu N., Tiwari R., Xu Y., & Yue L.Q. (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. https://doi.org/10.1080/10543406.2019.1657133

Wang X., Suttner L., Jemielita T., & Li X. (2022). Propensity score-integrated Bayesian prior approaches for augmented control designs: A simulation study. *Journal of Biopharmaceutical Statistics*, *32*(1), 170–190. https://doi.org/10.1080/10543406.2021.2011743

Yuan J., Liu J., Zhu R., Lu Y., & Palm U. (2019). Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. *Journal of Biopharmaceutical Statistics*, *29*(3), 558–573. https://doi.org/10.1080/10543406.2018.1559853

Zhang D., & Kim J. (2019). Use of propensity score and disease risk score for multiple treatments with time-to-event outcome: A simulation study. *Journal of Biopharmaceutical Statistics*, *29*(6), 1103–1115. https://doi.org/10.1080/10543406.2019.1584205