BASIC: A Bayesian Adaptive Synthetic Control Design for Phase II Clinical Trials

Journal Title
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Liyun Jiang[1,2], Peter F. Thall[2], Fangrong Yan[1], Scott Kopetz[3] and Ying Yuan[2]

Abstract

Background: Randomized controlled trials (RCTs) are considered the gold standard for evaluating experimental treatments, but often require large sample sizes. Single-arm trials require smaller sample sizes, but are subject to bias when using historical control data (HCD) for comparative inferences. This article presents a Bayesian adaptive synthetic control (BASIC) design that exploits HCD to create a hybrid of a single-arm trial and an RCT.

Methods: BASIC has two stages. In stage 1, a prespecified number of patients are enrolled in a single arm given the experimental treatment. Based on the stage 1 data, applying propensity score matching and Bayesian posterior prediction methods, the usefulness of the HCD for identifying a pseudo-sample of matched synthetic control patients for making comparative inferences is evaluated. If a sufficient number of synthetic controls can be identified, the single arm trial is continued. If not, the trial is switched to an RCT. The performance of BASIC is evaluated by computer simulation.

Results: BASIC achieves power and unbiasedness similar to an RCT, but on average requires a much smaller sample size, provided that the HCD patients are sufficiently comparable to the trial patients so that a good number of matched controls can be identified in the HCD. Compared to a single-arm trial, BASIC yields much higher power and much smaller bias.

Conclusions: The BASIC design provides a useful tool for exploiting HCD to improve the efficiency of single-arm phase II clinical trials, while addressing the problem of bias when comparing trial results to HCD. BASIC achieves power similar to an RCT, but may require a substantially smaller sample size.

Keywords

Real world data, historical data, augmented control, Bayesian adaptive design, randomized controlled trials

Background

The randomized controlled trial (RCT) is considered the gold standard for testing whether an experimental treatment, E, provides a therapeutic improvement over a standard control therapy, C. Lee and Feng 2005; Wieand 2005; Hariton and Locascio 2018 Randomization balances the E and C treatment arms with respect to both known and unknown patient characteristics, and provides unbiased estimators of causal E-versus-C treatment effects on clinical outcomes, such as response or survival time. Rubin 1978 An RCT may be challenging, however, due to the requirement of a large sample size. Consequently, many phase II oncology trials use single-arm designs and rely on historical control data (HCD) to compare the response rates of E and C and make go/no-go decisions for conducting a future phase III study. Simon 1989; Green and Dahlberg 1992; El-Maraghi and Eisenhauer 2008 Single-arm phase II trials require smaller sample sizes, but they may produce biased estimators of E-versus-C effects due to patient selection and systematic changes over time in patient prognosis or supportive care. The pervasive use of single-arm phase II trials has been cited as a major factor contributing to the high failure rate of phase III trials. Ratain and Sargent 2009; Tang et al. 2010; Rubinstein et al. 2009; Thall 2020 A comprehensive discussion of single-arm and randomized designs for phase II oncology trials is provided by Grayling et al (2019).Grayling et al 2019

There has been increasing interest in using HCD to bridge RCTs and single-arm trials, particularly for rare diseases or cancer subtypes. The US Food and Drug Administration (FDA) has released draft guidance on using HCD to support drug approval submission in 2019. FDA 2019 HCD may be obtained from one or more nonrandomized trials of C, randomized studies including C, electronic health records, medical claims and billing data, product and disease registration, or mobile health devices.

A well-established method for utilizing HCD to construct bias-corrected estimators of E-versus-C effects from non-randomized, observational data, including single-arm trials, is based on propensity score matching. Rosenbaum and Rubin 1983; Freemantle et al. 2013 Using their covariates, a patient from the HCD is selected to match each patient

[1] Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, Nanjing, China
[2] Department of Biostatistics,
[3] Department of Gastrointestinal Medical Oncology, University of Texas MD Anderson Cancer Center, TX, USA

Corresponding author:
Ying Yuan, Department of Biostatistics, University of Texas MD Anderson Cancer Center, TX, 77030, USA.
Email: yyuan@mdanderson.org

in the trial to have numerically similar propensity scores, which are estimated probabilities of receiving E. An attractive property of propensity scores is that, assuming no unobserved confounders, conditional on the propensity scores, the potential outcomes are independent of the actual treatments received, and the observed outcomes can be used to estimate the causal E-versus-C effect of interest. Rosenbaum and Rubin 1983 A matched pairs comparison addresses the problem that, without randomization, patients who received C may be systematically different from those who received E. For a large HCD sample, it may be possible to find several C patients who match each E patient, so 2-to-1 or 3-to-1 matching may be done to improve reliability. Rassen et al. 2012; Austin 2010 The dataset consisting of the matched control patients is referred to as synthetic controls, because they did not arise from randomization between E and C (Figure 1A).

Lin et al (2018) used propensity score matching methods to select additional patients from HCD to augment the active controls. Lin et al. 2018 Schmidli (2020) suggested using propensity score methods to utilize the HCD, rather than naive direct use of the HCD, in a single-arm trial.Schmidli 2020 Li et al (2020) provided a detailed discussion and practical considerations when using propensity score methods to incorporate HCD in clinical trials.Li et al. 2020 Thorlund et al (2020) provided a set of key questions to help researchers assess the validity and quality of trials utilizing synthetic control methodologies.Thorlund et al. 2020

Recently, propensity score matching has led to several new drug approvals by the FDA. For example, brineura (cerliponase alfa) was approved for treating a specific form of Batten disease based on a 22-patient single-arm trial compared to a control group with 42 patients. FDA 2017 Blinatumomab (Blincyto) was approved to treat Philadelphia chromosome-negative relapsed or refractory precursor B-cell acute lymphoblastic leukemia, based on a single-arm trial compared with a synthetic control sample constructed from 13 historical studies. Przepiorka et al. 2015 Ibrance (palbociclib) was approved by the FDA for treating men with HR+, HER2- metastatic breast cancer using synthetic control data. Pfizer 2019 Other methods also have been proposed for using HCD to design single-arm phase II trials. Thall and Simon 1990; Matano and Sambucini 2016 A review is given by Viele et al (2014). Viele et al. 2014

A limitation of propensity score matching is that the characteristics of patients treated with E may be so different from those of the HCD patients that very few matched pairs can be identified. The likelihood of this problem cannot be determined when designing a trial because the patient data for E are not yet available. In many cases, only when the trial is completed is it recognized that there are too few synthetic controls identified from HCD to provide an adequately powered comparison of E to C. This was the case for several early single-arm studies of the combination E = vemurafenib + irinotecan + cetuximab for BRAFV600E mutated colorectal cancer. The studies all enrolled patients with better prognosis, more indolent disease, better performance status, and longer prior survival, compared to HCD patients treated with C = irinotecan + cetuximab. Kopetz et al. 2021 Consequently, it was not possible to obtain a sufficient number of matched pairs to do a bias corrected comparison.

In this paper, rather than performing matched pairs estimation after the trial of E is completed, we propose a new Bayesian adaptive synthetic control (BASIC) phase II design that exploits HCD by doing pair matching during the trial. The BASIC design starts as a single-arm trial of E. During the trial, based on the HCD and interim data on E, the design predicts the number of HCD patients that can be matched to patients treated with E at the end of the trial. If this number is large enough to compare E to C with a prespecified power, the single-arm trial is continued. If not, the trial is switched to an RCT, with the randomization proportion chosen so that, at the end of the trial, the E and C sample sizes will be balanced. The BASIC design is illustrated in Figure 1B.

Gotte et al (2022) proposed an adaptive two-stage design including an interim decision to switch from a single-arm trial to a fixed-ratio RCT if a preference score that measures the comparability of covariates between patients receiving S and the HCD is lower than a fixed threshold. Gotte et al. 2022 There are two main differences between BASIC and this adaptive two-stage design. First, unlike the adaptive two-stage design, BASIC makes the interim decision by predicting the number of matched historical controls that will be obtained at the end of the trial. BASIC switches to an RCT if there are an insufficient number of predicted matched historical controls, which is a more direct approach than using a preference score. Because a preference score does not have an intuitive interpretation, choosing a numerical switching threshold is challenging. Gotte et al. (2022) used a threshold of 0.5, switching to an RCT if the preference score < 0.5. This may be problematic, for example, if the preference score < 0.5 but a sufficient number of matched controls can be found in the HCD, so there is no need to switch to an RCT. Second, when interim data satisfy the switching criterion, the adaptive two-stage design switches to a fixed-ratio RCT. In contrast, BASIC chooses the randomization ratio adaptively based on the effective sample size of the HCD, and thus is more flexible and more efficient, in that it only randomizes the number of patients needed to the control.

Methods

Propensity score Matching

A patient's propensity score Rosenbaum and Rubin 1983 is the probability of that patient receiving E, i.e., $e(X) = \Pr(\text{Treatment} = E|X)$, estimated based on the patient's baseline covariates X using a regression model, such as a logistic model, Austin 2011; Stuart 2010 fit to the data on E and C. The model includes all available patient baseline covariates that may be related to either the outcome or treatment, i.e., all potential confounders. A key property of propensity scores is that, if their distribution is balanced between the E and C samples, then all covariates used in the model also are balanced between the samples. Rosenbaum and Rubin 1983

Propensity scores can be used to identify synthetic controls by doing C-to-E patient matching, as follows. Haukoos and Lewis 2015 A patient treated with E is randomly selected, and a matched (synthetic) C patient then

is chosen so that their covariates give similar estimated propensity scores. The two patients' outcomes are recorded, the synthetic matched control is removed from the sample of C patients, and this is repeated until each E patient has a matched control. Using the sample of matched pairs, standard statistical methods can be used to estimate the mean E-versus-C effect and test whether it differs significantly from 0. Several propensity score matching algorithms are available. Austin 2011; Stuart 2010 We use nearest neighbor caliper propensity score matching with a caliper of 0.2 standard deviations, recommended by Rosenbaum and Rubin (1985) Rosenbaum and Rubin 1985, Austin (2011) Austin 2011 and others Stuart 2010; Caliendo and Kopeinig 2008, which can be implemented using the R packages MatchIt Ho et al. 2013 or Matching Sekhon 2011.

Bayesian adaptive synthetic control (BASIC) design

If only a small number of matched controls can be identified due to large differences between covariates of the HCD and E trial patients, then the synthetic matched control approach is not feasible, and it is better to conduct an RCT to obtain an unbiased comparison of E to C. The proposed BASIC design addresses this problem. Its key property is that, if interim data from a single-arm trial of E predict that an insufficient number of matched controls will be synthesized from the HCD, the design switches to an RCT between E and C.

A BASIC design allows multiple interim decisions, and can target an RCT with any randomization ratio, such as 2:1. For simplicity, we focus on a two-stage BASIC design with one interim look when half of the planned maximum of N patients have been accrued to the single-arm trial of E, and an RCT with a 1:1 randomization. The goal of this BASIC design is to emulate an RCT with $2N$ patients randomized fairly between E and C, with N selected based on a standard power calculation for the RCT (Figure 1B).

Predicting the number of matched controls N_s The BASIC design starts as a single-arm trial of E. At the interim decision, the trial data and the HCD are used to predict the number of matched controls, N_s, that can be identified after completing the single-arm trial. For patient i, let Y_i denote the binary or continuous outcome, $X_i = (1, x_{i1}, \cdots, x_{ir})$ denote the vector of r observed baseline patient covariates, and $T_i = 1$ if the patient is from the experimental treatment (E) arm and 0 if the patient is either a synthetic control from the HCD or a randomized control. Let N_h denote the sample size of the HCD.

Suppose that n patients are enrolled in the E arm at the interim decision. Based on the current observed data, $D_n = \{(Y_i, X_i, T_i), i = 1, \cdots, n + N_h\}$, we fit the logistic regression model

$$\text{logit}\{\Pr(T_i = 1|X_i)\} = X_i\boldsymbol{\eta},$$

where $\boldsymbol{\eta} = (\eta_0, \eta_1, \cdots, \eta_r)^T$ is a vector of model parameters. The estimated propensity score is

$$\hat{e}(X_i) = \frac{1}{1 + \exp(-X_i\hat{\boldsymbol{\eta}})}, \quad (1)$$

where $\hat{\boldsymbol{\eta}}$ denotes the estimator of $\boldsymbol{\eta}$.

We next apply Bayesian posterior prediction Gelman 2013 to predict the propensity scores of $N - n$ future patients to be enrolled in the E arm, and thereby the number of matched controls N_s, based on the observed interim data. The process for predicting N_s is as follows:

1. Under the following Bayesian model, compute the posterior distributions of the propensity scores.
   (a) Assume that the Logit transformation of propensity score $Z_i = \text{logit}(\hat{e}(X_i)) = X_i\hat{\boldsymbol{\eta}} \sim \text{Normal}(\mu, \sigma^2)$, which is generally reasonable by the central limit theorem;
   (b) Assume a noninformative prior for $\theta = (\mu, \sigma^2)$, e.g., $(\mu, \sigma^2) \sim (\sigma^2)^{-1}$;
   (c) Compute the posterior $\pi(\theta|Z_n)$, given the interim data $Z_n = \{Z_1, \cdots, Z_{n+N_h}\}$, resulting in $\mu \sim t\left(n-1, \hat{\mu}, \sqrt{\hat{\sigma}^2/n}\right)$ and $\sigma^2 \sim \text{Scale-inv-}\chi^2(n-1, \hat{\sigma}^2)$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and sample variance of Z in the E arm.
2. Simulate L propensity score datasets $\widetilde{\boldsymbol{e}}^{(1)}, \cdots, \widetilde{\boldsymbol{e}}^{(L)}$, each corresponding to $N - n$ future patients, from the posterior predictive propensity score distribution, as follows. For each $\ell = 1, \cdots L$,
   (a) Simulate $\theta^{(\ell)}$ from $\pi(\theta|Z_n)$.
   (b) Simulate $\widetilde{Z}^{(\ell)} = \left(\widetilde{z}_1^{(\ell)}, \cdots, \widetilde{z}_{N-n}^{(\ell)}\right)$ from $f\left(Z|\theta^{(\ell)}\right)$, and compute the $N - n$ propensity scores $\widetilde{\boldsymbol{e}}^{(\boldsymbol{\ell})} = \text{expit}\{\widetilde{Z}^{(\ell)}\}$.
3. Predict N_s based on propensity score matching, as follows. For each $\ell = 1, \cdots L$, given the predicted propensity score dataset $\widetilde{\boldsymbol{e}}^{(\boldsymbol{\ell})}$ for $N - n$ future patients, and the estimated propensity scores of n enrolled patients and N_h historical patients, use the propensity score matching algorithm to identify historical patients matched to the N patients in the E arm. Let $N_s^{(\ell)}$ denote the total number of matched historical control patients. Predict N_s using the qth percentile of $\{N_s^{(1)}, \cdots, N_s^{(L)}\}$. We recommend using the 50th percentile, i.e. the median ($q = 0.5$).

Interim decision The usefulness of the HCD is quantified by the synthesis efficiency, SynEff $= N_s/N$, the predicted proportion of controls that can be synthesized from the HCD. SynEff $= 1$ if N matched controls can be synthesized from the HCD, whereas SynEff $= 0$ if no matched controls can be synthesized. Given a specified fixed threshold π between 0 and 1, if SynEff $< \pi$ then it is considered unlikely that the HCD will provide N synthetic controls, so the BASIC design switches to an RCT. If SynEff $> \pi$ then the single-arm trial of E with N patients is continued, at the end up to N matched controls are identified, and a pair-matched estimator of the E-versus-C effect is computed.

The interim decision rule of BASIC is as follows:

(i) If SynEff $< \pi$, switch to an RCT in which $2N - n - N_s$ future patients are enrolled during stage 2 and randomized between E and C in the ratio $(N - n) : (N - N_s)$.

(ii) If SynEff $\geq \pi$, continue the single-arm trial of E during stage 2 and enroll $N - n$ additional patients.

The randomization ratio is chosen to obtain approximately N patients in each of the E and C arms at the end of the trial. The stage 2 randomization produces a hybrid control arm, consisting of N_s synthetic nonrandomized control patients and $N - N_s$ concurrent randomized control patients.

The threshold π that controls whether the trial is switched to an RCT may be chosen by conducting preliminary computer simulations of the trial using several different values, based on the design's operating characteristics, including power and type I error rate. A practical approach is to choose π to give power within a prespecified margin (e.g., 5%) of a targeted power (e.g., 80%). A value of π between 0.8 and 0.9 typically yields good operating characteristics. If switching to an RCT is not logistically feasible, setting $\pi = 0$ gives a conventional single-arm trial with synthetic controls constructed from the HCD at the end of the trial. If desired, one can also add a "discard" rule: If SynEff $< \pi_l$, discard the HCD and switch to an RCT with the randomization ratio $(N - n) : N$, where π_l is a small value (e.g., 0.05). The rationale for this rule is that if the HCD differs substantially from the trial population in that at most a few patients can be matched, one may completely discard the HCD and take the simpler and cleaner approach of only using the trial data to compare E to C.

After completing stage 2 of the BASIC design, the logistic propensity score model is updated by fitting it to the final trial data and the HCD, and propensity score matching is used to construct a final set of synthetic controls. Depending on the interim decision, this can be either a fully synthetic control arm, or a hybrid control arm as described above. A standard frequentist test (e.g., a t-test or chi-square test) or statistical estimates (e.g., means, proportions, or regression model-based estimates, with confidence intervals) can be used to evaluate the E-versus-C treatment effect. Austin 2011; Stuart 2010; Caliendo and Kopeinig 2008 Alternatively, Bayesian posterior probabilities and credible intervals can be used for final inferences. While the randomization and adaptive interim decisions may give final sample sizes of synthetic or hybrid controls not exactly equal to N, this deviation typically is small and has a negligible impact on the operating characteristics of the design, as shown in the simulation study given below.

Interim futility stopping

If desired, the following Bayesian interim futility stopping rule may be added: If Pr(treatment effect of $E >$ treatment effect of C + targeted improvement | data) $< \lambda$, then stop the trial for futility, where λ is a fixed cutoff chosen by preliminary simulations (e.g., $\lambda = 0.1$). This futility stopping rule is evaluated based on the interim observed E data and matched HCD. The results are reported in Appendix I-II in the Supplementary Material. Other than the futility stopping rule, if appropriate, any other type of interim rules (e.g., sample size re-calculation) also can be added to BASIC to fit trial objectives.

Results

Simulation settings

We evaluated the operating characteristics of the BASIC design by computer simulation, including comparisons to three alternative designs: an RCT with 1: 1 randomization to E and C; a conventional single-arm design with HCD used as a comparator without matching; and a single-arm design with synthetic controls generated at the end of the trial using propensity score matching. The synthetic-control design is a special case of BASIC with $\pi = 0$. We considered both a binary endpoint (e.g., response) and a continuous endpoint (e.g., biomarker level), with four covariates, including two binary confounders X_1, X_2 and two continuous confounders X_3, X_4. We assumed that the HCD includes 160 patients, and simulated their baseline covariates from a mixed population including patients both similar and dissimilar to the E patients. This was done as follows:

1) The covariate data of n_h comparable historical patients were generated from a joint distribution. Specifically, we first simulated (X_1, X_2, X_3, X_4) from a multivariate normal distribution MVN(μ_1, Σ_1) with $\mu_1 = (0, 0, 0, 0)$, the diagonal of Σ_1 being $(1, 1, 0.25^2, 0.25^2)$ and the off-diagonal elements being 0.1. We then converted X_1 and X_2 to binary covariates using the cut point 0, such that mean values of X_1 and X_2 were 0.5.
2) The covariate data of $N_h - n_h$ non-comparable patients were generated from a different joint distribution. Specifically, we first simulated (X_1, X_2, X_3, X_4) from a multivariate normal distribution MVN(μ_0, Σ_0) with $\mu_0 = (0, 0, 0.8, 1.5)$, the diagonal of Σ_0 being $(1, 1, 0.25^2, 0.5^2)$ and the off-diagonal elements being 0.1. We then converted X_1 and X_2 to binary covariates using cut points $\Phi^{-1}(0.2)$ and $\Phi^{-1}(0.8)$ such that mean values of X_1 and X_2 were 0.2 and 0.8, respectively, where $\Phi^{-1}(\cdot)$ is the quantile function of a standard normal random variable.

In each simulated trial, we controlled the synthesis efficiency: SynEff $= N_s/N$ at a fixed value by setting the value of n_h. For the binary endpoint, the outcomes Y_i were generated from the logit model

$$\text{logit}\ \{\Pr(Y_i = 1 \mid T_i, X_i)\} = \beta T_i + \sum_{k=1}^{4} \alpha_k X_{ik}. \quad (2)$$

For the continuous endpoint, the outcomes were generated from the normal linear model

$$Y_i = \beta T_i + \sum_{k=1}^{4} \alpha_k X_{ik} + \epsilon_i, \quad (3)$$

where $\epsilon_i \sim iid\ N(0, 1)$. We assumed confounder effects $\alpha_1 = 0.12, \alpha_2 = -2.6, \alpha_3 = -0.96, \alpha_4 = 2$ for binary endpoints, and $\alpha_1 = 0.9, \alpha_2 = 0.4, \alpha_3 = 1.2, \alpha_4 = -0.2$ for continuous endpoints. We simulated trials with a planned sample size of $N = 80$ per arm for the RCT, and 80% power to detect an improvement (treatment effect size) δ of 0.19 (i.e., $\beta = 1.21$) for the response probabilities of E

versus C, or 0.41 (i.e., $\beta = 0.45$) for the standardized difference between the means of a continuous endpoint. For the conventional single-arm design, we used design parameters estimated from the HCD, e.g., historical response rate, and an improvement δ, to estimate the sample size, obtain 80% power, and control type I error at 5%.

We considered the values SynEff $= 1.0, 0.8, 0.5, 0.3, 0.1$ and 0 to represent a wide range of degrees of usefulness of the HCD to allow matched controls to be synthesized. We considered BASIC designs with one interim decision based on $\pi = 0.9$ after $n = 40$ of the $N = 80$ patients per arm were enrolled. For all designs, at the end of the trial a one-sided Z-test for binomial proportions or t-test for continuous endpoints was used to test the null hypothesis of no E-versus-C effect versus the alternative that E provides an improvement, with a significance level 0.05. We simulated 5000 trials using each design in each simulation scenario, and calculated the type I error rate, power, average total sample size, and relative bias $\frac{|\hat{\delta}-\delta|}{\delta}$, where $\hat{\delta}$ is the estimate of the effect size. Figures 2 and 3 show the simulation results for binary and continuous endpoints, respectively. Detailed simulation results are shown in Tables A1 and A2 of Appendix II in the Supplementary Material.

Simulation results

Figure 2 illustrates the simulation results for binary endpoints. As expected, the RCT yields high power, low bias, and a type I error rate near 0.05, but it requires the largest sample size. The single-arm design requires the smallest sample size, but has by far the lowest power (Fig 2B) and largest bias (Fig 2C) of all four designs, especially when patients in the HCD differ substantially from those in the trial (i.e. SynEff $= 0, 0.1$ or 0.3). The single-arm design also fails to control the type I error rate at the nominal level, with values substantially lower than 0.05 (Fig 2A). Because it is based on synthetic controls matched to the E patients, the synthetic-control design has much higher power and much lower bias than the single-arm design. In the case where an insufficient number of controls can be synthesized due to large differences between HCD patients and trial patients, the synthetic-control design has lower power and higher bias than the RCT.

BASIC has the best overall performance among all four designs. Compared to the RCT, BASIC has similar power, bias, and type I error, but requires a substantially smaller sample size (Fig 2D). For example, when matched controls for all E patients can be synthesized from the HCD (i.e., SynEff $= 1$ in Fig 2D), the sample size of BASIC is about half that of the RCT. BASIC has much higher power and much smaller bias than the conventional single-arm design. Because BASIC adaptively determines whether there is a need to randomize patients to C, depending on the usefulness of the HCD, BASIC avoids the loss of power seen with the synthetic-control design when an inadequate number of controls can be synthesized from the historical data (i.e., SynEff $= 0.1, 0.3$ in Fig2 B). When synthesis efficiency = 1, BASIC has slightly higher power than a RCT. This is because, in this case, each patient in the trial has a matched control and the propensity score matching often results in better covariate balance than complete randomization, thus leading to higher power than an RCT. This phenomenon also was reported by Joffe (1999) Joffe 1999 and Ali et al (2019) Ali et al. 2019. For the same reason, the type I error of BASIC is slightly lower than the nominal value.

In summary, BASIC solves the problem of bias with the single-arm design, and solves the problem of low power with the synthetic-control design when the HCD patients have characteristics different from those of trial patients.

Figure 3 shows the simulation results of the four designs with a continuous endpoint. These results are qualitatively very similar to those seen for a binary endpoint. BASIC again has the best overall performance, with power and bias similar to the RCT, but substantially smaller sample size.

Sensitivity analysis

We also studied the sensitivity of the BASIC design to (1) the time point used for the interim analysis, (2) sample size of 40 per arm, (3) effects of unmeasured confounders not included in the patients covariates, and (4) patient drift. Figures A1-A6 in Appendix I in Supplementary Material show the simulation results, and detailed results are shown in Tables A3-A8 in Appendix II in Supplementary Material.

We also considered cases with the interim analysis at an earlier time point, when $t = 20\%$ patients are enrolled, and a later time point when $t = 90\%$ patients are enrolled. As shown in Figures A1 and A2, type I error, power and bias are generally similar to the case with $t = 50\%$, suggesting that BASIC is robust to the choice of interim time. Of note, the sample size is slightly sensitive to the interim time. For example, if SynEff $= 0.8$, when more data are used to estimate propensity scores (e.g., $t = 90\%$), BASIC requires a smaller sample size than cases where fewer data values are used (e.g., $t = 20\%$ or 50%). This probably is because BASIC can estimate propensity scores more accurately by using more data. In general, we recommend $t = 50\%$, which provides enough data for an interim decision, but also is early enough to allow the adaptation to be effective.

Figures A3 and A4 show simulation results for binary and continuous endpoints, respectively, when the sample size is 40 for the treatment arm, and 80 for the historical data. The sample size of the single-arm design is estimated based on HCD, as described before. The treatment effect size is adjusted to ensure that the RCT yields 80% power. The results are generally similar to what was reported previously, showing that BASIC still has the comparative advantages seen with larger sample sizes.

To assess how the four methods behave when some covariates that affect treatment or outcome are not observed, i.e. are unmeasured confounders, we considered the case of a binary endpoint with sample sizes $N = 80$ per arm for the RCT, 80 for the E arm of the synthetic-control design and BASIC, and $N_h = 160$ for the HCD. For the single-arm design, we estimated the sample size based on the HCD. The methods for generating covariates were similar to those in the earlier simulations. For patients in the trial of E, we considered $\mu_1 = (0, 0, 0, 0, 0)$, the diagonal of Σ_1 as $(1, 1, 0.25^2, 0.25^2, 0.25^2)$ and the off-diagonal elements 0.1, and a cut point 0 used to convert X_1 and X_2 to binary covariates, to simulate the covariates. For the historical C patients, we generated covariates from a mixture distribution: 1) covariates of n_h comparable

historical C patients were generated from the same distributions used for E patients, 2) covariates of $N_h - n_h$ non-comparable patients were generated from different distributions, with $\mu_0 = (0, 0, 0.8, 1.5, 1)$, the diagonal of Σ_0 as $(1, 1, 0.25^2, 0.5^2, 0.25^2)$ and the off-diagonal elements all 0.1, and cut points $\Phi^{-1}(0.2)$ and $\Phi^{-1}(0.8)$ used to convert X_1 and X_2 to binary covariates, respectively. We considered BASIC deigns with the values SynEff $= 1, 0.8, 0.5, 0.3, 0.1$ and 0, obtained by setting the value of n_h. The outcomes were generated from the logit model

$$\text{logit}\left\{\Pr\left(Y_i = 1 | T_i, X_i\right)\right\} = \beta T_i + \sum_{k=1}^{5} \alpha_k X_{ik}, \quad (4)$$

with $\alpha_1 = -0.5, \alpha_2 = -1.5, \alpha_3 = -2, \alpha_4 = -0.5, \alpha_5 = 4$, and $\beta = 1.02$ (under which RCT detects an effect size $\delta = 0.2$ with the power of 80%). We assumed that the covariate X_5 was not observed and not included in the propensity model. Figure A5 shows the simulation results. In the presence of unmeasured confounders, BASIC yields satisfactory performance similar to RCT, but with smaller sample size, and higher power and lower bias than the single-arm and synthetic-control designs. Compared to the ideal case where all confounders are included in the propensity model, the power of BASIC is slightly lower with slightly higher bias. This highlights the importance of including all potential confounders in the propensity model Rosenbaum and Rubin 1983; Freemantle et al. 2013, if they are available.

In some trials, the baseline patient covariate distribution may drift over time and become different between stages I and II. To evaluate the performance of BASIC in the presence of drift, we considered the case of a binary endpoint with four covariates, X_1, X_2, X_3, X_4, and an interim analysis when $t = 50\%$ patients are enrolled. The covariate distribution and data generation procedure of the E patients before the interim analysis were the same as those in the simulations. For patients after the interim analysis, the mean value of X_4 drifted higher by 0.2 standard deviation. The treatment effect size was adjusted accordingly, so that the RCT has 80% power. Figure A6 shows the simulation results. In the presence of this population shift, BASIC still had the best overall performance, with power and bias similar to the RCT, but smaller sample size; and higher power and lower bias than the single-arm and synthetic-control designs. BASIC is robust to patient drift because matching with the HCD largely eliminates the impact of patient drift.

BASIC with interim futility stopping

We also investigated the performance of BASIC with the Bayesian interim futility stopping rule Pr(treatment effect of $E >$ treatment effect of $C \mid$ data) $< \lambda$, where the cutoff λ was calibrated to control the probability of early stopping in the case where E provides an improvement δ over C at 10%. We considered binary and continuous endpoints, and treatment effect size δ = 0.20 (i.e., $\beta = 1.25$) for the response probabilities of E versus C, or 0.41 (i.e., $\beta = 0.46$) for the standardized difference between the means of a continuous endpoint. The remaining settings and data generation procedure were the same as those described in the simulations. Figures A7-A8 in Appendix I and Tables A9-A10 in Appendix II in Supplementary Material show the simulation results. In general, BASIC yields power and bias similar to an RCT, but with smaller sample size; higher power and lower bias than the single-arm design; and higher power than the synthetic-control design. If desired, a standard frequentist-based approach could be used for interim futility stopping.

Conclusions

We have proposed a new hybrid phase II design, BASIC, that exploits HCD to do approximately unbiased estimation of E-versus-C effects similarly to an RCT. The key property of BASIC is that, depending on the usefulness of the HCD to allow synthetic controls to be identified, it may adaptively switch from a single-arm trial to an RCT. Our simulations show that BASIC (1) avoids the problem of biased estimation when single-arm trial results are compared to HCD, (2) is superior to the common approach of doing a comparison based on synthetic matched controls identified at the end of a single-arm trial, and (3) performs similarly to an RCT in terms of power and bias, but with a much smaller sample size.

We have focused on the case that starts with a single-arm trial and then may adaptively switch to an RCT. The BASIC design can be modified to start as an RCT and then adaptively adjust the randomization ratio or switch to a single-arm trial (the extreme case with randomization probability 0 to the control) based on the predicted number of controls that can be synthesized from the HCD. While we estimated propensity scores using a logistic regression model, a nonparametric approach can be used to improve robustness of the propensity score estmation. For example, one may use generalized boosted models McCaffrey et al. 2004, which can estimate a nonlinear relationship between covariates and propensity scores. BASIC relies on estimated propensity scores to predict the expected number of matched patients at the end of the trial, and uses this to decide whether to switch to an RCT. The interim decision time should be chosen appropriately, so that there are a reasonable number of interim data values to reliably fit and estimate the propensity model. The interim decision time should be chosen and calibrated by simulation, while accounting for other clinical and logistic considerations.

As with all propensity score-based methods, the validity of BASIC relies on the assumption that there are no unmeasured confounders. Consequently, when building the propensity model, it is critical to include as many key prognostic factors as feasible in the model, based on clinical judgement and historical data. Ali et al (2019) summarized methods for dealing with unmeasured confounders. Ali et al. 2019 Because the assumption of no unmeasured confounders cannot be tested, a sensitivity analysis provides a useful tool to assess the potential impact if this assumption is violated. Rosenbaum 1991

The interim adaptation by BASIC makes it challenging to implement blinding if the trial is switched to an RCT. This might be done by establishing an independent data safety monitoring committee and a coordinating center to perform the interim analysis and decisions, and the possible randomization in stage 2. More generally, FDA guidance

provides useful recommendations to maintain the integrity of trials that use adaptive designs. FDA 2019

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Lee JJ, Feng L: Randomized phase II designs in cancer clinical trials: current status and future directions. J Clin Oncol 23(19): 4450-4457, 2005.

Wieand HS. Randomized phase II trials: what does randomization gain? J Clin Oncol 23(9):1794-5, 2005.

Hariton E, Locascio JJ: Randomised controlled trialsthe gold standard for effectiveness research. BJOG 125(13):1716, 2018.

Rubin DB: Bayesian inference for causal effects: the role of randomization. Ann Stat 6(1):34-58, 1978.

Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10(1):1-10, 1989.

Green SJ, Dahlberg S: Planned versus attained design in phase II clinical trials. Stat Med 11(7):853-862, 1992.

El-Maraghi RH, Eisenhauer EA: Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. J Clin Oncol 26(8):1346-1354, 2008.

Ratain MJ, Sargent DJ: Optimising the design of phase II oncology trials: the importance of randomisation. Eur J Cancer 45(2):275-280, 2009.

Tang H, Foster NR, Grothey A, et al: Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. J Clin Oncol 28(11):1936-1941, 2010.

Rubinstein L, Crowley J, Ivy P, et al: Randomized phase II designs. Clin Cancer Res 15(6):1883-90, 2009.

Thall PF: Statistical Remedies for Medical Researchers. Springer Series in Pharmaceutical Statistics, 2020.

Grayling MJ, Dimairo M, Mander AP, et al: A review of perspectives on the use of randomization in phase II oncology trials. JNCI: J Natl Cancer Inst 111(12):1255-1262, 2019.

US Food and Drug Administration: Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics guidance for industry. May 2019. Accessed October 30, 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/ submitting-documents-using-real-worlddata-and-real-world-evidence-fda-drugsand-biologics-guidance

Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41-55, 1983.

Freemantle N, Marston L, Walters K, et al: Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. BMJ 347: f6409, 2013.

Lin J, GamaloSiebers M, Tiwari R: Propensity score matched augmented controls in randomized clinical trials: A case study. Pharm Stat 17(5):629-647, 2018.

Schmidli H, Hring DA, Thomas M, et al: Beyond randomized clinical trials: Use of external controls. Clin Pharmacol Ther 107(4):806-816, 2020.

Li Q, Lin J, Chi A, et al: Practical considerations of utilizing propensity score methods in clinical development using real-world and historical data. Contemp Clin Trials 97:106123, 2020.

Rassen JA, Shelat AA, Myers J, et al: Onetomany propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf 21(S2):69-80, 2012.

Austin PC: Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. Am J Epidemiol 172(9):1092-1097, 2010.

Thorlund K, Dron L, Park JJ, et al: Synthetic and external controls in clinical trialsa primer for researchers. Clin Epidemiol 12:457-467, 2020.

US Food and Drug Administration: FDA approves first treatment for a form of batten disease [press release]. Online, April 2017.

Przepiorka D, Ko CW, Deisseroth A, et al: FDA approval: blinatumomab. Clin Cancer Res 21(18):4035-4039, 2015.

Pfizer: US Food and Drug Administration approves Ibrance (Palbociclib) for the treatment of men with HR+, HER2-Metastatic breast cancer [www.pfizer.Com]. Available at 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, 2019.

Thall PF, Simon R: Incorporating historical control data in planning phase II clinical trials. Stat Med 9(3):215-228, 1990.

Matano F, Sambucini V: Accounting for uncertainty in the historical response rate of the standard treatment in singlearm twostage designs based on Bayesian power functions. Pharm Stat 15(6):517-530, 2016.

Viele K, Berry S, Neuenschwander B, et al: Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat 13(1):41-54, 2014.

Kopetz S, Guthrie KA, Morris VK, et al: Randomized trial of irinotecan and cetuximab with or without vemurafenib in BRAF-mutant metastatic colorectal cancer (SWOG S1406). J Clin Oncol 39(4):285-294, 2021.

Gotte H, Kirchner M, Krisam J, et al: An adaptive design for early clinical development including interim decision for singlearm trial with external controls or randomized trial. Pharm Stat 21(3):625640, 2022.

Austin PC: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46(3):399-424, 2011.

Stuart EA: Matching methods for causal inference: A review and a look forward. Stat Sci 25(1):1-21, 2010.

Haukoos JS, Lewis RJ: The propensity score. JAMA 314(15):1637-1638, 2015.

Rosenbaum PR, Rubin DB: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33-38, 1985.

Caliendo M, Kopeinig S: Some practical guidance for the implementation of propensity score matching. J Econ Surv 22(1):31-72, 2008.

Ho DE, Imai K, King G, et al: MatchIt: Nonparametric preprocessing for parametric causal inference. Software for using matching methods in R. Available at http://gking.harvard.edu/matchit/, 2013.

Sekhon JS: Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw 42(7):1-52, 2011.

Gelman A, Carlin J B, Stern H S, et al: Bayesian data analysis. CRC press, 2013.

Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 150(4):327-333, 1999.

Ali MS, Prieto-Alhambra D, Lopes LC, et al: Propensity score methods in health technology assessment: principles, extended applications, and recent advances. Front Pharmacol 10:973, 2019.

McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 9(4): 403-425, 2004.

Rosenbaum PR: Sensitivity analysis for matched case-control studies. Biometrics 47(1):87-100, 1991.

US Food and Drug Administration: Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. November 2019.

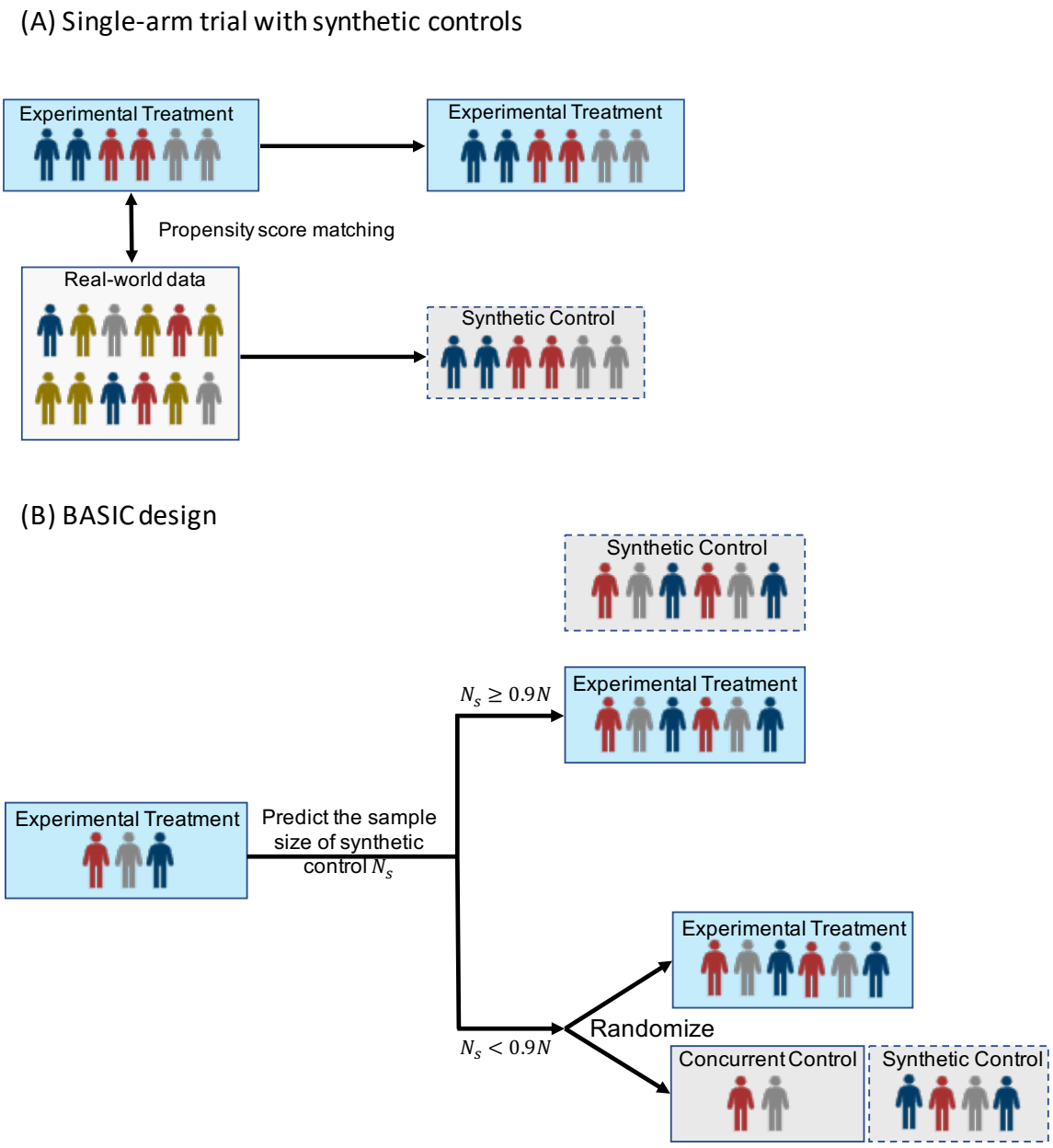

Figure 1. (A) Single-arm trial with synthetic controls obtained by matching from historical control data, used in a final comparative analysis (synthetic control design) to emulate an RCT; (B) Schema of the BASIC design emulating an RCT with $2N$ patients randomized equally between the E and C arms. In stage 1, n (e.g., $N/2$) patients are enrolled to a single E arm. At the interim decision, the number of controls N_s that can be synthesized from the HCD by the completion of the trial is predicted. If $N_s > 0.9N$ (a sufficient number of controls can be synthesized), the trial is continued as a single-arm trial of E using synthetic controls; otherwise, the trial is switched to an RCT, resulting in a hybrid control sample consisting of both concurrent randomized controls and synthetic controls.

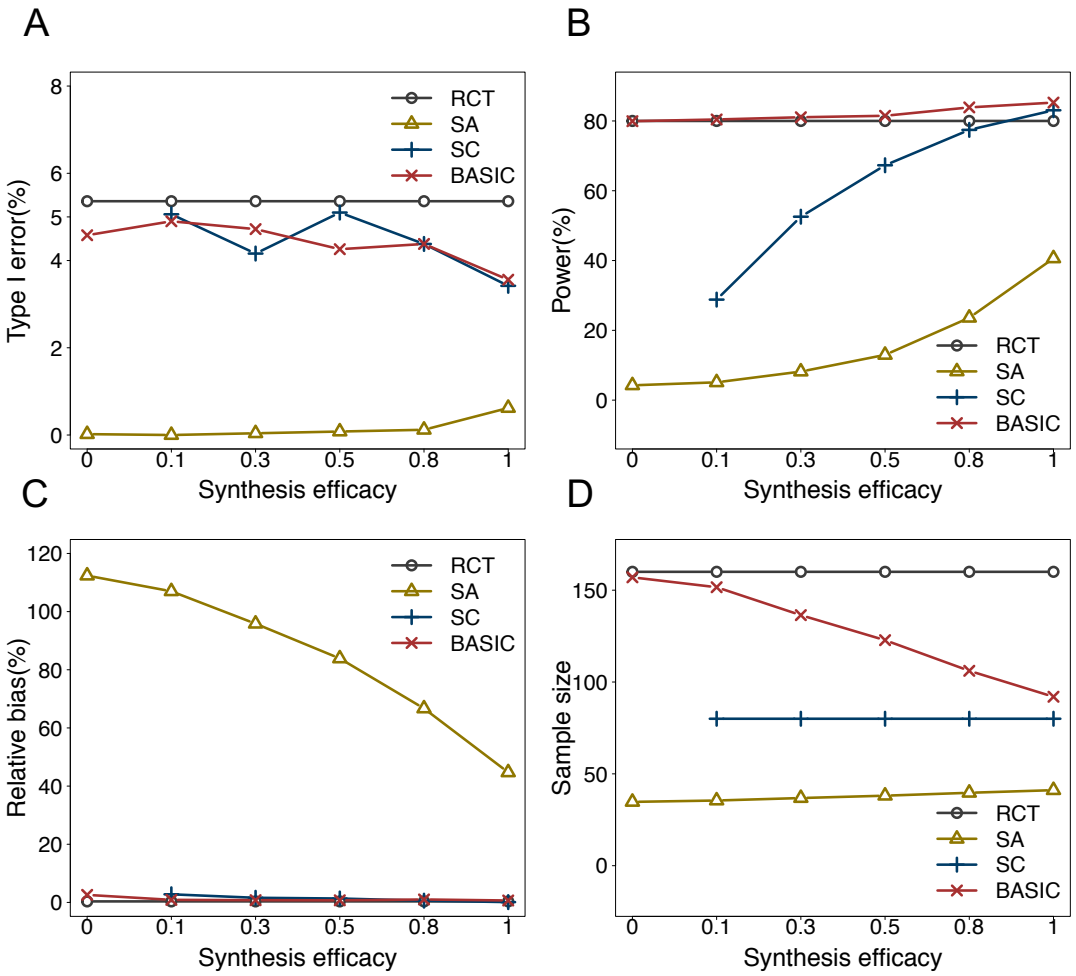

Figure 2. Simulation results of the RCT, single-arm design (SA), single-arm design with synthetic controls (SC) and BASIC design, for a binary endpoint under different synthesis efficiencies from the historical control data. If SynEff $= 0$, i.e., no historical controls are chosen, SC becomes infeasible and thus its results are null values.

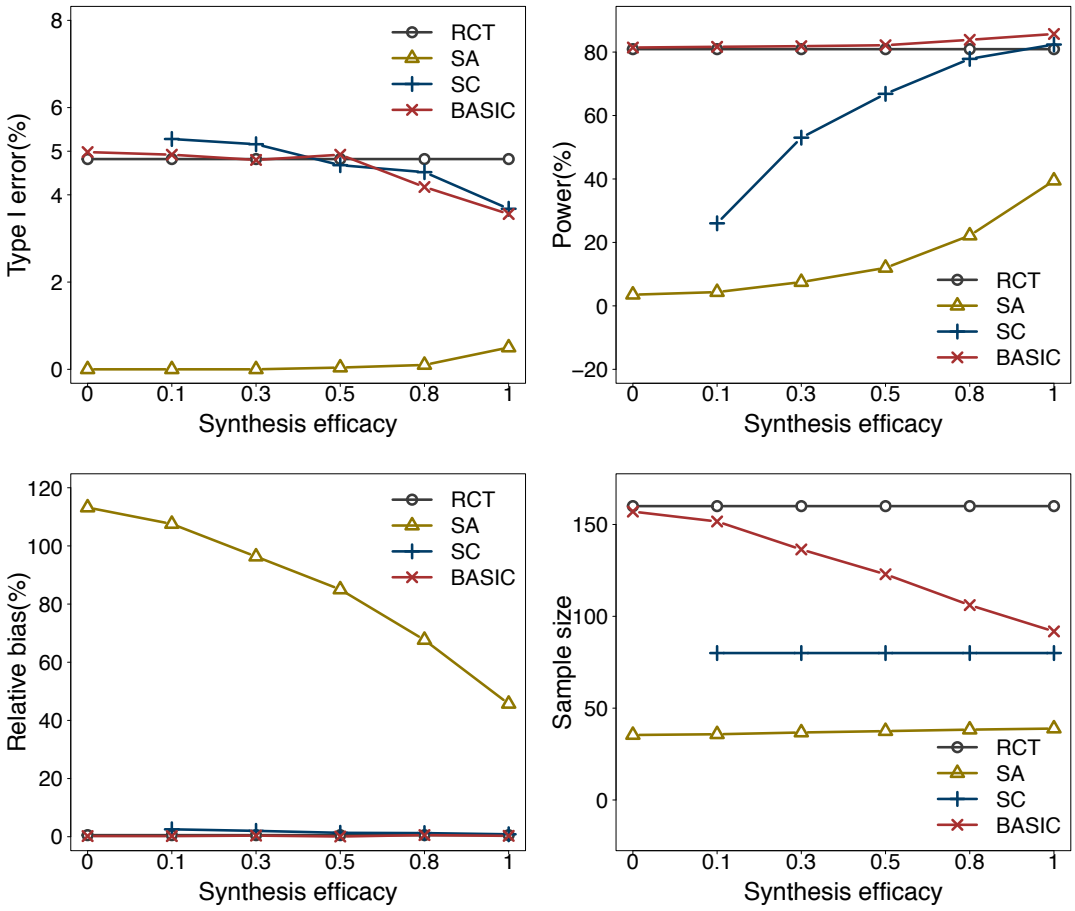

Figure 3. Simulation results of the RCT, single-arm design (SA), single-arm design with synthetic control (SC) and BASIC design for a continuous endpoint under different synthesis efficiencies from the historical control data. If SynEff $= 0$, i.e., no historical controls are chosen, SC becomes infeasible and thus its results are null values.